Il Teorema del Reverendo Bayes

Il teorema di Bayes rappresenta una delle più profonde intuizioni nella storia del pensiero matematico e scientifico. Non si tratta semplicemente di una formula, ma di un principio fondamentale che ha rivoluzionato il modo in cui concepiamo la conoscenza, l'incertezza e il ragionamento razionale. Questa equazione apparentemente semplice ha aperto prospettive completamente nuove in campi che spaziano dalla filosofia alla fisica quantistica, dalla medicina all'intelligenza artificiale, dimostrando una portata universale che pochi altri risultati matematici possono vantare.

La forza dirompente del teorema risiede nella sua capacità di invertire la direzione del ragionamento causale. Tradizionalmente, la probabilità procede dalle cause agli effetti: conosciamo la probabilità che una causa produca un certo effetto. Il teorema di Bayes ci permette di procedere in direzione opposta: osservati certi effetti, quali sono le probabilità delle diverse cause che potrebbero averli prodotti? Questa inversione è determinante perché nella pratica scientifica e nella vita quotidiana raramente osserviamo le cause direttamente, ma solo i loro effetti. Dobbiamo inferire, dedurre, ricostruire le cause a partire dalle conseguenze osservabili.

Consideriamo un esempio medico che illustra perfettamente questa inversione. Un medico sa che una certa malattia produce determinati sintomi con una certa probabilità. Questa è la direzione naturale del ragionamento causale: dalla malattia ai sintomi. Ma il problema diagnostico reale è l'inverso: osservati certi sintomi in un paziente, qual è la probabilità che il paziente abbia quella specifica malattia? Il teorema di Bayes fornisce il meccanismo matematico preciso per operare questa inversione, trasformando la conoscenza delle probabilità causa-effetto nella conoscenza delle probabilità effetto-causa.

Thomas Bayes e il suo Tempo

Thomas Bayes nacque probabilmente nel 1701 a Londra, in un periodo di straordinario fermento intellettuale in Inghilterra. Era figlio di Joshua Bayes, uno dei primi sei ministri presbiteriani ordinati dopo che il Toleration Act del 1689 aveva concesso una limitata libertà religiosa ai dissidenti protestanti. Questa appartenenza alla comunità dei non conformisti, che rifiutavano l'autorità della Chiesa d'Inghilterra, avrebbe segnato profondamente la vita di Thomas, escludendolo dalle università di Oxford e Cambridge, riservate agli anglicani, ma al contempo inserendolo in una tradizione intellettuale che valorizzava il pensiero indipendente e critico.

Non si hanno certezze sulla formazione di Bayes, ma è probabile che abbia studiato privatamente, forse presso l'accademia dissenziente di Moorfields a Londra, istituzioni educative alternative create dai non conformisti per aggirare l'esclusione dalle università tradizionali. Queste accademie, paradossalmente, erano spesso all'avanguardia nell'insegnamento delle scienze naturali e della matematica, meno vincolate dalla tradizione scolastica rispetto alle università ufficiali.

Bayes seguì le orme del padre nel ministero religioso e nel 1731 divenne ministro della cappella di Mount Sion a Tunbridge Wells, nel Kent, una cittadina termale frequentata dall'aristocrazia londinese. Questa posizione relativamente tranquilla gli permise di coltivare i suoi interessi matematici. Nel 1742 fu eletto Fellow della Royal Society, il massimo riconoscimento scientifico britannico dell'epoca, probabilmente sulla base di un'opera non sopravvissuta sulla matematica del calcolo infinitesimale o per la sua difesa dei principi del calcolo newtoniano contro le critiche del filosofo George Berkeley.

Il Contesto Matematico del XVIII Secolo

La Royal Society del XVIII secolo era un'istituzione straordinaria, un crocevia dove si incontravano scienziati naturali, matematici, filosofi e persino teologi interessati alle questioni scientifiche. Le riunioni della società vedevano discussioni animate sui più recenti sviluppi in astronomia, fisica, matematica e filosofia naturale. In questo ambiente, Bayes venne certamente a contatto con le questioni più avanzate della matematica del suo tempo, inclusi i problemi emergenti nella teoria della probabilità.

Il XVIII secolo era infatti un periodo cruciale per lo sviluppo della probabilità come disciplina matematica. I lavori pionieristici di Pascal e Fermat sulla probabilità nei giochi d'azzardo, risalenti alla metà del XVII secolo, avevano posto le basi concettuali. Christiaan Huygens aveva pubblicato nel 1657 il primo trattato sistematico sulla probabilità, "De Ratiociniis in Ludo Aleae". Jakob Bernoulli aveva dimostrato la legge dei grandi numeri nella sua opera postuma "Ars Conjectandi" del 1713, stabilendo rigorosamente che le frequenze relative convergono alle probabilità teoriche quando il numero di prove aumenta indefinitamente.

Abraham de Moivre, ugonotto francese rifugiato in Inghilterra e contemporaneo di Bayes, aveva pubblicato "The Doctrine of Chances" nel 1718, applicando la matematica del calcolo infinitesimale ai problemi probabilistici e sviluppando la teoria delle aspettative matematiche. De Moivre aveva anche scoperto l'approssimazione normale della distribuzione binomiale, un risultato fondamentale che avrebbe anticipato il teorema del limite centrale.

In questo contesto intellettuale vivace, il problema che avrebbe condotto Bayes al suo teorema era naturale ed urgente. Tutti i lavori precedenti sulla probabilità procedevano in una direzione: data una situazione con probabilità note (un dado equo, un'urna con composizione nota), quali sono le probabilità dei diversi risultati possibili? Ma il problema inverso rimaneva aperto e concettualmente problematico: osservati certi risultati, cosa possiamo inferire sulle probabilità che li hanno generati? Come possiamo risalire dagli effetti alle cause?

Il Manoscritto Postumo e il Contributo di Richard Price

Thomas Bayes morì nel 1761, e la sua opera più importante non fu mai pubblicata durante la sua vita. Il manoscritto, intitolato "An Essay towards solving a Problem in the Doctrine of Chances", rimase tra le sue carte personali. Fu Richard Price, matematico, filosofo morale e amico di Bayes, a scoprire questo lavoro e a comprenderne l'importanza rivoluzionaria.

Price stesso era una figura intellettuale di grande rilievo. Nato nel Galles nel 1723, era diventato un influente predicatore dissenziente a Londra e aveva sviluppato profondi

interessi in matematica, filosofia morale ed economia. Era amico di Benjamin Franklin e avrebbe successivamente sostenuto pubblicamente sia la Rivoluzione Americana che quella Francese, attirando su di sé violente critiche dai conservatori britannici. La sua opera principale "A Review of the Principal Questions in Morals" del 1758 aveva tentato di fondare l'etica su basi razionali.

Price capì che il manoscritto di Bayes conteneva una soluzione genuinamente nuova a un problema fondamentale. Il 23 dicembre 1763, presentò il lavoro alla Royal Society, accompagnandolo con una lettera introduttiva in cui spiegava il contesto e l'importanza del risultato. Price scrisse che Bayes aveva trovato "una regola attraverso la quale possiamo giudicare riguardo la probabilità che un evento abbia certe cause, quando abbiamo conoscenza della frequenza con cui è accaduto in precedenza."

L'articolo fu pubblicato nelle Philosophical Transactions della Royal Society nel 1763, due anni dopo la morte di Bayes. Il problema specifico che Bayes considerava era il seguente: supponiamo di osservare una sequenza di successi e fallimenti in esperimenti ripetuti (come lanci di una moneta), senza conoscere la vera probabilità di successo. Basandosi su queste osservazioni, cosa possiamo dire sulla probabilità sottostante?

L'Esperimento Mentale di Bayes

Bayes affrontò il problema in modo particolarmente ingegnoso, usando quello che oggi chiameremmo un esperimento mentale geometrico. Immaginò una tavola perfettamente levigata su cui una palla (chiamiamola W) viene lanciata e si ferma in una posizione casuale. Questa prima palla determina una linea che divide la tavola in due regioni. Poi vengono lanciate n altre palle, e si registra quante cadono alla sinistra della linea (chiamiamole "successi") e quante alla destra ("fallimenti").

Il problema è: osservando il numero di successi e fallimenti, cosa possiamo inferire sulla posizione della prima palla W, che determina la probabilità di successo? Se W è molto a sinistra, la probabilità di successo è bassa; se è molto a destra, è alta. Bayes sviluppò un metodo per calcolare, dato il numero osservato di successi e fallimenti, la distribuzione di probabilità per la posizione di W.

La soluzione di Bayes coinvolge quello che oggi chiamiamo integrazione bayesiana. Egli assunse inizialmente che tutte le posizioni per W fossero equiprobabili (quello che oggi chiamiamo prior uniforme), e poi calcolò come questa distribuzione dovesse essere aggiornata alla luce delle osservazioni. Il risultato finale, espresso in termini moderni, è che la distribuzione posteriore per la probabilità di successo p, dati m successi in n prove, è una distribuzione Beta con parametri (m+1, n-m+1).

La densità posteriore è proporzionale a p^m × (1-p)^(n-m), che è massima quando p = m/n, la frequenza osservata dei successi. Ma crucialmente, la distribuzione posteriore ci dice molto di più della semplice stima puntuale: ci dice quanto siamo incerti su p, come questa incertezza dipende dal numero di osservazioni, e ci permette di calcolare intervalli di credibilità per p.

Le Applicazioni di Price

Richard Price comprese che questo metodo aveva implicazioni vastissime. Nel suo saggio che accompagnava il lavoro di Bayes, Price applicò il metodo a problemi di interesse pratico, incluso quello demografico di stimare la popolazione totale di un paese basandosi su campionamenti parziali. Questo era un problema di grande importanza all'epoca, quando i censimenti sistematici non esistevano ancora e la dimensione della popolazione era cruciale per questioni di tassazione e potere militare.

Price vedeva anche implicazioni teologiche nel lavoro di Bayes. In una sezione del suo saggio, sostenne che il metodo bayesiano forniva supporto razionale all'argomento del disegno divino: l'ordine osservato nell'universo rende estremamente probabile l'esistenza di un progettista intelligente. Sebbene questo argomento teologico sia oggi considerato problematico, rivela come Price vedesse il teorema di Bayes come uno strumento generale per il ragionamento induttivo in condizioni di incertezza.

Pierre-Simon Laplace: Sistemazione e Generalizzazione

Se Bayes fu il pioniere che aprì la strada, Pierre-Simon Laplace fu l'architetto che costruì l'edificio maestoso dell'inferenza bayesiana. Laplace, nato in Normandia nel 1749, fu uno dei più grandi matematici e scienziati della storia, con contributi fondamentali in meccanica celeste, equazioni differenziali, teoria del potenziale e, naturalmente, teoria della probabilità.

Laplace riscoprì indipendentemente il teorema di Bayes intorno al 1774, apparentemente senza conoscere il lavoro di Bayes. Nella sua "Mémoire sur la probabilité des causes par les événements" presentata all'Académie Royale des Sciences di Parigi, Laplace enunciò chiaramente il principio di inversione probabilistica e lo applicò a diversi problemi pratici. La formulazione di Laplace era più generale e matematicamente più rigorosa di quella di Bayes.

La grande opera di Laplace sulla probabilità fu la monumentale "Théorie Analytique des Probabilités", pubblicata per la prima volta nel 1812 e ripubblicata in edizioni successive sempre ampliate. Quest'opera sistematica conteneva tutto il sapere matematico sulla probabilità accumulato fino ad allora, presentato con rigore analitico e potenti tecniche di calcolo. Laplace vi sviluppò estensivamente quello che oggi chiamiamo inferenza bayesiana, applicandola a innumerevoli problemi in astronomia, geodesia, demografia e scienze fisiche.

Il Principio di Indifferenza

Uno dei contributi concettuali più importanti di Laplace fu l'enunciazione esplicita del principio di ragione insufficiente, che egli chiamò principio di indifferenza. Questo principio afferma che, in assenza di qualsiasi informazione che favorisca un'ipotesi rispetto a un'altra, tutte le ipotesi dovrebbero essere considerate equiprobabili a priori. Per Laplace, questo era un principio razionale necessario per rendere oggettiva l'inferenza probabilistica.

Laplace applicò sistematicamente il suo metodo bayesiano ai grandi problemi scientifici dell'epoca. In astronomia, lo usò per stimare le orbite dei pianeti e delle comete basandosi su osservazioni imprecise. In geodesia, lo applicò al problema di determinare la forma della Terra da misurazioni di archi di meridiano. In demografia, seguì Price nell'usare metodi bayesiani per stimare popolazioni e tassi di natalità e mortalità.

Un'applicazione particolarmente famosa del metodo di Laplace riguardava il rapporto tra nascite maschili e femminili. Era stato osservato empiricamente che nascevano leggermente più maschi che femmine, ma c'era dibattito su se questa differenza fosse reale o dovuta a fluttuazioni casuali. Laplace usò l'inferenza bayesiana applicata ai registri delle nascite a Parigi e Londra per concludere con alta probabilità che il fenomeno era reale e stimò con precisione il rapporto vero.

L'Approssimazione di Laplace

Laplace sviluppò anche quella che oggi chiamiamo l'approssimazione di Laplace, un metodo per approssimare integrali complessi che appaiono frequentemente nel calcolo di probabilità posteriori. Quando il posteriore ha un picco pronunciato (situazione tipica quando si hanno molti dati), Laplace dimostrò che può essere approssimato con una distribuzione normale centrata sul massimo posteriore, con varianza determinata dalla curvatura del logaritmo del posteriore in quel punto. Questa approssimazione è ancora ampiamente usata oggi quando l'integrazione esatta è intrattabile.

Un aspetto filosofico importante del lavoro di Laplace fu la sua interpretazione deterministica della probabilità. Nel suo famoso "Essai philosophique sur les probabilités" del 1814, Laplace espresse la sua visione deterministica dell'universo: un'intelligenza sufficientemente vasta che conoscesse le posizioni e velocità di tutte le particelle dell'universo potrebbe, usando le leggi della meccanica, calcolare l'intero passato e futuro. Per Laplace, la probabilità nasceva solo dall'ignoranza umana, non da qualche indeterminismo intrinseco nella natura. Questa prospettiva sarebbe stata radicalmente scossa nel XX secolo dalla meccanica quantistica.

Il Declino e l'Oblio nel XIX Secolo

Nonostante l'enorme prestigio di Laplace, il suo approccio bayesiano alla probabilità gradualmente cadde in disfavore durante il XIX secolo. Diverse ragioni contribuirono a questo declino. Prima di tutto, emersero crescenti preoccupazioni sulla legittimità e l'oggettività del principio di indifferenza. Il matematico francese Joseph Bertrand mostrò con i suoi famosi paradossi che il principio di indifferenza poteva portare a contraddizioni quando applicato ingenuamente.

Il paradosso più famoso di Bertrand riguarda una corda casuale su un cerchio. La domanda è: qual è la probabilità che una corda scelta casualmente su un cerchio sia più lunga del lato di un triangolo equilatero inscritto nel cerchio? Bertrand mostrò che, a seconda di come si interpreta "casualmente", si ottengono risposte diverse: 1/2, 1/3, o 1/4. Questo sollevò seri dubbi sulla presunta oggettività delle probabilità a priori uniformi.

Inoltre, il XIX secolo vide l'emergere di una concezione più empirista e positivista della scienza. L'idea che si dovessero incorporare "credenze pregresse" nell'analisi scientifica appariva a molti come una contaminazione soggettiva del metodo scientifico. La scienza doveva basarsi solo sui dati oggettivi, non su assunzioni a priori che potevano riflettere pregiudizi personali o metafisici.

Questo clima intellettuale favorì lo sviluppo di approcci alternativi alla statistica. La scuola frequentista, che sarebbe diventata dominante nel XX secolo, cominciò a prendere forma. Questa scuola interpreta la probabilità esclusivamente in termini di frequenze a lungo termine in esperimenti ripetibili. Un evento ha probabilità p se, ripetendo l'esperimento un numero infinito di volte, la frequenza relativa con cui l'evento si verifica converge a p.

Il Trionfo del Frequentismo: Fisher, Neyman, Pearson

Il XX secolo vide l'affermazione definitiva della statistica frequentista come paradigma dominante, guidata da tre figure titaniche: Ronald Fisher, Jerzy Neyman e Egon Pearson. Questi statistici svilupparono un framework completo per l'inferenza statistica che deliberatamente escludeva ogni riferimento a probabilità a priori o a credenze soggettive.

Ronald Fisher, forse il più grande statistico del XX secolo, sviluppò negli anni '20 e '30 il concetto di verosimiglianza (likelihood) e il metodo della massima verosimiglianza. Per Fisher, l'inferenza statistica doveva basarsi solo sulla funzione di verosimiglianza, che esprime quanto bene diversi valori del parametro spiegano i dati osservati. Fisher respinse esplicitamente l'uso di distribuzioni a priori, che considerava soggettive e arbitrarie. La sua opera "Statistical Methods for Research Workers" del 1925 divenne il testo fondamentale della statistica moderna e plasmò intere generazioni di statistici.

Fisher introdusse anche il concetto di test di significatività e il p-value, strumenti che sarebbero diventati onnipresenti nella ricerca scientifica. L'idea è di calcolare, sotto l'assunzione che l'ipotesi nulla sia vera, la probabilità di osservare dati estremi quanto o più di quelli effettivamente osservati. Un p-value molto piccolo è considerato evidenza contro l'ipotesi nulla. Questo approccio non richiede specificare probabilità a priori e appariva quindi più oggettivo.

Jerzy Neyman e Egon Pearson svilupparono negli anni '30 una teoria rigorosa dei test di ipotesi basata sulla teoria delle decisioni. Loro concettualizzarono il test statistico come una regola di decisione: dati certi dati, decidiamo se accettare o rigettare un'ipotesi. Questa decisione può essere errata in due modi: rigettare l'ipotesi quando è vera (errore di tipo I) o accettarla quando è falsa (errore di tipo II). Neyman e Pearson svilupparono una teoria matematica elegante per trovare test ottimali che minimizzassero queste probabilità di errore.

L'approccio di Neyman-Pearson aveva il grande vantaggio di fornire garanzie di performance a lungo termine. Un test con livello di significatività α del 5% garantisce che, se l'ipotesi nulla è vera e ripetiamo l'esperimento molte volte, rigettereremo erroneamente l'ipotesi al massimo nel 5% dei casi. Queste garanzie di frequenza non richiedono assunzioni su probabilità a priori e apparivano quindi più solide e oggettive.

Il frequentismo divenne il paradigma ortodosso nella statistica per gran parte del XX secolo. I textbook standard insegnavano metodi frequentisti quasi esclusivamente. Le principali riviste statistiche pubblicavano principalmente ricerca frequentista. Gli statistici bayesiani divennero una minoranza emarginata, spesso visti come eretici o eccentrici.

Il Revival Bayesiano: De Finetti, Savage, Lindley

Tuttavia, alcuni pensatori visionari mantennero viva la tradizione bayesiana e la svilupparono in direzioni nuove e profonde. Il revival bayesiano del XX secolo fu guidato principalmente da tre figure: Bruno de Finetti, Leonard Jimmie Savage e Dennis Lindley.

Bruno de Finetti, matematico e probabilista italiano, rivoluzionò completamente la fondazione concettuale della probabilità bayesiana. Nella sua opera monumentale "La prévision: ses lois logiques, ses sources subjectives" del 1937 e in lavori successivi, de Finetti sviluppò un'interpretazione radicalmente soggettivista della probabilità. Per de Finetti, la probabilità non esiste come proprietà oggettiva del mondo esterno, ma è semplicemente una misura del grado di credenza personale di un individuo.

La genialità di de Finetti fu di mostrare che questa interpretazione soggettivista non porta all'anarchia o all'arbitrarietà totale. Egli dimostrò che qualsiasi sistema di credenze probabilistiche che voglia essere coerente deve obbedire agli assiomi della probabilità. La coerenza è definita operazionalmente: un agente è coerente se non può essere "dutch booked", cioè indotto ad accettare una serie di scommesse che garantiscono una perdita certa indipendentemente da come si risolvano gli eventi incerti.

De Finetti dimostrò anche il suo famoso teorema di rappresentazione: una sequenza infinita di variabili aleatorie considerate scambiabili (cioè tali che la loro distribuzione congiunta è invariante rispetto a permutazioni) può essere rappresentata come condizionalmente indipendente e identicamente distribuita dato un parametro aleatorio. Questo risultato fornisce una giustificazione profonda per l'approccio bayesiano: trattare parametri come variabili aleatorie non è un'artificio matematico ma una conseguenza logica dell'assunzione naturale di scambiabilità.

Leonard Jimmie Savage, statistico americano, sviluppò negli anni '50 una teoria assiomatica della decisione in condizioni di incertezza che forniva fondamenta rigorose all'inferenza bayesiana. Nel suo libro fondamentale "The Foundations of Statistics" del 1954, Savage mostrò che qualsiasi agente razionale che soddisfi certi assiomi plausibili di comportamento decisionale deve comportarsi come se stesse massimizzando l'utilità attesa rispetto a una distribuzione di probabilità soggettiva che viene aggiornata secondo il teorema di Bayes quando arrivano nuove informazioni.

Gli assiomi di Savage includono principi come: se preferisci l'azione A all'azione B quando sei sicuro che si verificherà l'evento E, allora devi preferire A a B anche quando E è solo probabile (principio della cosa certa). Savage mostrò che questi assiomi, sebbene individualmente molto ragionevoli, implicano collettivamente che l'agente deve avere una distribuzione di probabilità ben definita su tutti gli eventi e deve aggiornarla bayesianamente. Questo fornì una giustificazione normativa potente per l'approccio bayesiano: non è solo un metodo statistico tra tanti, ma l'unico metodo coerente con principi razionali di decisione.

Dennis Lindley, statistico britannico, fu probabilmente il più influente divulgatore e difensore del bayesianesimo nella seconda metà del XX secolo. Inizialmente formato nella tradizione frequentista di Fisher, Lindley si convertì al bayesianesimo e divenne il suo campione più eloquente. Nei suoi numerosi libri e articoli, incluso il classico "Introduction to Probability and Statistics from a Bayesian Viewpoint" del 1965, Lindley articolò chiaramente i vantaggi concettuali dell'approccio bayesiano.

Lindley enfatizzò particolarmente come l'inferenza bayesiana fornisca risposte dirette alle domande che gli scienziati realmente vogliono porre. Un ricercatore tipicamente vuole sapere: "Dati questi dati, qual è la probabilità che la mia ipotesi sia vera?" L'approccio bayesiano risponde direttamente a questa domanda calcolando P(ipotesi|dati). L'approccio frequentista, al contrario, calcola P(dati|ipotesi), la probabilità dei dati assumendo che l'ipotesi sia vera, che è la risposta a una domanda diversa.

Lindley identificò anche molti paradossi e incoerenze nei metodi frequentisti. Il più famoso è il paradosso di Lindley stesso: per campioni molto grandi, un test frequentista può rigettare fortemente un'ipotesi nulla (p-value minuscolo) mentre l'analisi bayesiana fornisce alta probabilità posteriore a favore dell'ipotesi nulla. Questo accade perché i metodi frequentisti e bayesiani rispondono a domande fondamentalmente diverse.

La Rivoluzione Computazionale

Nonostante questi profondi contributi concettuali, il bayesianesimo rimaneva in gran parte un'esercitazione teoretica fino agli anni '90. Il problema era computazionale: calcolare distribuzioni posteriori per modelli realistici richiede l'integrazione su spazi ad alta dimensionalità, un compito praticamente impossibile con i metodi analitici tradizionali o con i computer disponibili fino agli anni '80.

La svolta arrivò con lo sviluppo e la diffusione dei metodi Markov Chain Monte Carlo (MCMC), particolarmente l'algoritmo di Gibbs sampling e il suo software di implementazione BUGS (Bayesian inference Using Gibbs Sampling). Sebbene le idee di base dell'MCMC fossero note sin dagli anni '50, grazie ai lavori pionieristici di Nicholas Metropolis e dei suoi collaboratori a Los Alamos, la loro applicazione sistematica all'inferenza bayesiana avvenne solo negli anni '90.

L'idea fondamentale dell'MCMC è ingegnosa: invece di calcolare il posteriore analiticamente, costruiamo una catena di Markov (una sequenza di stati casuali dove ogni stato dipende solo dallo stato precedente) la cui distribuzione limite è esattamente il posteriore che vogliamo calcolare. Dopo un periodo iniziale di "riscaldamento", la catena produce campioni dalla distribuzione posteriore. Questi campioni possono poi essere usati per approssimare qualsiasi quantità di interesse: medie posteriori, intervalli di credibilità, probabilità posteriori di ipotesi.

Il software BUGS, sviluppato da un gruppo di ricercatori a Cambridge guidato da David Spiegelhalter, rese l'MCMC accessibile a ricercatori non specialisti. Un utente poteva specificare il suo modello in un linguaggio semplice, e BUGS automaticamente costruiva e faceva girare la catena di Markov appropriata. Questo abbassò drasticamente la barriera d'ingresso all'analisi bayesiana.

Negli anni 2000 e 2010, nuovi algoritmi ancora più potenti furono sviluppati. L'Hamiltonian Monte Carlo (HMC), che usa idee dalla fisica hamiltoniana per proporre mosse più efficienti nello spazio dei parametri, si rivelò particolarmente efficace per modelli ad alta dimensionalità. Il No-U-Turn Sampler (NUTS), un'estensione dell'HMC che adatta automaticamente i suoi parametri, fu implementato nel software Stan, che rapidamente divenne lo standard de facto per l'inferenza bayesiana seria.

Questi sviluppi computazionali, combinati con la crescita esponenziale della potenza di calcolo disponibile, trasformarono completamente il paesaggio. Modelli bayesiani che sarebbero stati completamente intrattabili nel 1980 divennero routine negli anni 2000. Gerarchie complesse con decine o centinaia di parametri, modelli non lineari sofisticati, processi stocastici elaborati – tutti divennero analizzabili con metodi bayesiani.

Questo ha portato a un'esplosione di applicazioni bayesiane in ogni campo scientifico. In epidemiologia, modelli bayesiani gerarchici permettono di stimare tassi di malattia specifici per piccole aree geografiche condividendo informazione tra aree. In genetica, metodi bayesiani sono essenziali per ricostruire alberi filogenetici e mappare loci associati a malattie complesse. In astrofisica, l'inferenza bayesiana è usata per caratterizzare exoplaneti e stimare parametri cosmologici. In machine learning, approcci bayesiani forniscono principi per regolarizzazione, selezione di modelli e quantificazione dell'incertezza.

Oggi, il dibattito frequentismo-bayesianesimo, pur non completamente risolto, ha perso molto della sua asprezza. Molti statistici adottano un approccio pragmatico, usando metodi bayesiani quando sono appropriati e computazionalmente fattibili, e metodi frequentisti in altre situazioni. C'è anche crescente riconoscimento che i due approcci possono essere complementari piuttosto che contraddittori. Il teorema di Bayes, da curiosità matematica del XVIII secolo, è diventato uno strumento centrale nel toolkit della scienza moderna.

Lo Spazio di Probabilità e la Misura

Per comprendere profondamente il teorema di Bayes, dobbiamo iniziare dalle fondamenta rigorose della teoria della probabilità moderna, stabilite da Andrey Kolmogorov nel suo monumentale lavoro "Grundbegriffe der Wahrscheinlichkeitsrechnung" del 1933. Kolmogorov comprese che la probabilità poteva essere formulata come un caso speciale della teoria della misura, fornendo finalmente basi matematiche solide a una disciplina che per secoli aveva vissuto in uno stato di ambiguità concettuale.

Uno spazio di probabilità è una terna (Ω, F, P) dove:

- Ω è lo spazio campionario (l'insieme di tutti i possibili esiti di un esperimento aleatorio)
- F è una σ-algebra di sottoinsiemi di Ω (gli eventi di cui ha senso parlare di probabilità)
- P è una misura di probabilità su F

La σ-algebra F deve soddisfare proprietà di chiusura:

- Deve contenere l'insieme vuoto $\varnothing$ e l'insieme intero Ω
- Deve essere chiusa rispetto al complemento (se $A \in F$, anche $A^c \in F$)
- Deve essere chiusa rispetto a unioni numerabili (se $A_1, A_2, A_3, \ldots \in F$, anche $\bigcup_i A_i \in F$)

Queste proprietà garantiscono che possiamo eseguire le operazioni logiche naturali sugli eventi.

La misura di probabilità P assegna a ogni evento A in F un numero reale P(A) soddisfacendo tre assiomi fondamentali:

Primo assioma - Non-negatività: P(A) ≥ 0 per ogni evento A. Questo cattura l'intuizione che le probabilità non possono essere negative.

Secondo assioma - Normalizzazione: P(Ω) = 1. Questo dice che qualcosa deve certamente accadere – la probabilità dell'intero spazio campionario è uno.

Terzo assioma - Additività numerabile (σ-additività): Se A_1, A_2, A_3, ... è una sequenza di eventi mutuamente disgiunti (nessun esito può appartenere a più di uno di essi), allora:

$P(\cup_{i=1}^{\infty} A_i) = \Sigma_{i=1}^{\infty} P(A_i)$

Quest'ultimo assioma merita particolare attenzione perché è più forte di quanto potrebbe sembrare inizialmente. Richiede che l'additività valga non solo per unioni finite, ma per unioni di infiniti numerabili eventi. Questa proprietà è cruciale per garantire continuità e permettere l'uso del calcolo infinitesimale nella teoria della probabilità. Senza σ-additività, molti risultati fondamentali, incluse le leggi dei grandi numeri e il teorema del limite centrale, non potrebbero essere dimostrati rigorosamente.

Da questi tre assiomi possiamo derivare tutte le proprietà familiari della probabilità. Ad esempio, possiamo dimostrare che P(∅) = 0 segue necessariamente: l'insieme vuoto può essere scritto come unione infinita di se stesso, quindi $P(\varnothing) = \Sigma_{i=1}^{\infty} P(\varnothing)$, che è possibile solo se P(∅) = 0. Possiamo dimostrare che la probabilità del complemento di A è P(A^c) = 1 - P(A), perché A e A^c sono disgiunti e la loro unione è Ω che ha probabilità uno. Possiamo dimostrare il principio di inclusione-esclusione per calcolare la probabilità di unioni di eventi che si sovrappongono.

La Definizione Formale di Probabilità Condizionata

La probabilità condizionata è il concetto chiave che ci permette di aggiornare le nostre credenze alla luce di nuove informazioni. Data un'evento B con probabilità positiva, la probabilità condizionata di un evento A dato B è definita come:

P(A|B) = P(A ∩ B) / P(B)

assumendo P(B) > 0.

Questa definizione cattura formalmente un'intuizione profonda. Quando sappiamo che B è avvenuto, lo spazio degli esiti possibili si restringe da Ω a B. Dobbiamo ri-normalizzare le probabilità in questo spazio ristretto. Gli esiti in cui avviene sia A che B mantengono le loro probabilità relative, ma vengono scalate dal fattore 1/P(B) in modo che il nuovo spazio condizionato abbia probabilità totale uno.

È essenziale verificare che P(·|B), vista come funzione del suo primo argomento, costituisca essa stessa una misura di probabilità sullo spazio (Ω, F). Dobbiamo verificare i tre assiomi:

1. **Non-negatività:** Poiché sia $P(A \cap B)$ che $P(B)$ sono non-negativi e $P(B)$ è positivo, il rapporto è non-negativo.

2. **Normalizzazione:** $P(\Omega|B) = P(\Omega \cap B)/P(B) = P(B)/P(B) = 1$

3. **Additività numerabile:** Se A_1, A_2, ... sono mutuamente disgiunti, allora $(A_1 \cap B)$, $(A_2 \cap B)$, ... sono anch'essi mutuamente disgiunti, e usando l'additività di P otteniamo: $P(\cup_i A_i \mid B) = \Sigma_i P(A_i|B)$

Questa verifica non è pedanteria matematica ma ha conseguenze importanti. Significa che tutto il macchinario della teoria della probabilità può essere applicato a probabilità condizionate. Possiamo parlare di probabilità condizionate iterate, di variabili aleatorie condizionate, di aspettazioni condizionate, e così via. La struttura matematica è ricorsiva e auto-simile.

Derivazione Rigorosa del Teorema di Bayes

Dalla definizione di probabilità condizionata, possiamo derivare il teorema di Bayes con pochi passi di algebra. Cominciamo scrivendo la definizione in due modi differenti.

Per primo, esprimiamo la probabilità condizionata di A dato B:

$P(A|B) = P(A \cap B) / P(B)$

Moltiplicando entrambi i lati per P(B) otteniamo:

$P(A \cap B) = P(A|B) \cdot P(B)$

Ora esprimiamo la probabilità condizionata di B dato A:

$P(B|A) = P(B \cap A) / P(A)$

Moltiplicando entrambi i lati per P(A):

$P(B \cap A) = P(B|A) \cdot P(A)$

Ma l'intersezione è un'operazione commutativa: $A \cap B = B \cap A$. Questi sono esattamente gli stessi eventi, solo descritti in ordine diverso. Quindi i lati sinistri delle due equazioni sono identici, il che significa che i lati destri devono essere uguali:

$P(A|B) \cdot P(B) = P(B|A) \cdot P(A)$

Dividendo entrambi i lati per P(B), assumendo che $P(B) \neq 0$, otteniamo la forma classica del **teorema di Bayes**:

$P(A|B) = [P(B|A) \cdot P(A)] / P(B)$

Questa è una delle formule più importanti in tutta la matematica. Nonostante la sua semplicità, le sue conseguenze sono rivoluzionarie. Ci permette di invertire il

condizionamento: se conosciamo P(B|A), possiamo calcolare P(A|B). Possiamo passare dalla probabilità degli effetti date le cause alla probabilità delle cause dati gli effetti.

Interpretazione Semantica dei Termini

Ogni termine nella formula di Bayes ha un significato specifico e un nome convenzionale nella letteratura statistica. Consideriamo il caso tipico dove A rappresenta un'ipotesi o teoria e B rappresenta dati o evidenza osservata.

- **P(A|B)** è chiamata **probabilità posteriore**. È la probabilità dell'ipotesi A dopo aver osservato l'evidenza B. Questo è tipicamente ciò che vogliamo calcolare – quanto dovremmo credere nell'ipotesi alla luce dei nuovi dati?

- **P(B|A)** è chiamata **verosimiglianza** (likelihood). È la probabilità di osservare l'evidenza B se l'ipotesi A fosse vera. La verosimiglianza codifica quanto bene l'ipotesi predice o spiega i dati osservati.

- **P(A)** è chiamata **probabilità a priori** o semplicemente **prior**. È la probabilità dell'ipotesi prima di osservare l'evidenza B. Rappresenta ciò che credevamo (o dovremmo credere) sull'ipotesi sulla base di informazione pregressa.

- **P(B)** è chiamata **evidenza marginale** o **costante di normalizzazione**. È la probabilità totale di osservare l'evidenza B, mediata su tutte le possibili ipotesi.

Possiamo riscrivere il teorema di Bayes in forma mnemonica che evidenzia queste interpretazioni:

Posteriore = (Verosimiglianza × Prior) / Evidenza

O in termini ancora più intuitivi: la nostra credenza aggiornata in un'ipotesi è proporzionale a quanto bene l'ipotesi spiega i dati moltiplicato per quanto credevamo nell'ipotesi a priori.

La Regola della Probabilità Totale e la Forma Estesa

Spesso la probabilità marginale P(B) nel denominatore non è immediatamente disponibile. Dobbiamo calcolarla usando tutte le informazioni disponibili. Qui entra in gioco la regola della probabilità totale, che è essenziale per rendere il teorema di Bayes operativo in situazioni pratiche.

Supponiamo che abbiamo una collezione di ipotesi mutuamente esclusive ed esaustive A_1, A_2, ..., A_n. "Mutuamente esclusive" significa che al massimo una di esse può essere vera: $A_i \cap A_j = \varnothing$ per $i \neq j$. "Esaustive" significa che almeno una di esse deve essere vera: $A_1 \cup A_2 \cup ... \cup A_n = \Omega$. In terminologia matematica, $\{A_1, A_2, ..., A_n\}$ forma una partizione dello spazio campionario.

Qualsiasi evento B può essere scomposto secondo questa partizione:

$B = B \cap \Omega = B \cap (A_1 \cup A_2 \cup ... \cup A_n) = (B \cap A_1) \cup (B \cap A_2) \cup ... \cup (B \cap A_n)$

Poiché gli A_i sono mutuamente disgiunti, anche le intersezioni $(B \cap A_i)$ lo sono. Quindi per l'additività della probabilità:

$P(B) = P(B \cap A_1) + P(B \cap A_2) + ... + P(B \cap A_n)$

Ma dalla definizione di probabilità condizionata, $P(B \cap A_i) = P(B|A_i) \cdot P(A_i)$. Sostituendo:

$P(B) = \Sigma_{i=1}^{n} P(B|A_i) \cdot P(A_i)$

Questa è la **regola della probabilità totale**. Sostituendo questo nel teorema di Bayes, otteniamo la **forma estesa**:

$P(A_n|B) = [P(B|A_n) \cdot P(A_n)] / [\Sigma_{i=1}^{n} P(B|A_i) \cdot P(A_i)]$

Questa forma è estremamente utile nella pratica. Ci dice che per calcolare la probabilità posteriore di una specifica ipotesi A_n, dobbiamo:

1. Calcolare il numeratore: la verosimiglianza dei dati sotto quell'ipotesi moltiplicata per il prior di quell'ipotesi
2. Calcolare il denominatore: sommare il prodotto di verosimiglianza per prior su tutte le ipotesi possibili
3. Dividere il primo per il secondo

Il denominatore serve come fattore di normalizzazione, garantendo che le probabilità posteriori di tutte le ipotesi sommino a uno, come deve essere per una distribuzione di probabilità appropriata.

Estensione al Caso Continuo

Quando lavoriamo con variabili aleatorie continue piuttosto che discrete, il teorema di Bayes assume una forma analoga ma usando densità di probabilità invece di probabilità puntuali.

Consideriamo un parametro continuo θ (che può essere scalare o vettoriale) e dati osservati x. Invece di probabilità discrete, abbiamo funzioni di densità:

- $p(\theta)$: densità a priori che descrive la nostra credenza sul parametro prima di osservare i dati
- $p(x|\theta)$: funzione di verosimiglianza che descrive la densità di probabilità dei dati condizionata sul valore del parametro
- $p(\theta|x)$: densità a posteriori, ciò che vogliamo calcolare

Il teorema di Bayes nel caso continuo si scrive:

$p(\theta|x) = [p(x|\theta) \cdot p(\theta)] / p(x)$

dove la densità marginale dei dati è calcolata mediante integrazione:

$p(x) = \int p(x|\theta) \cdot p(\theta)\, d\theta$

Se θ è multidimensionale, questo è un integrale multiplo su tutto lo spazio dei parametri. Per esempio, se $\theta = (\theta_1, \theta_2)$ è bidimensionale:

$$p(x) = \iint p(x|\theta_1, \theta_2) \cdot p(\theta_1, \theta_2) \, d\theta_1 \, d\theta_2$$

Calcolare questi integrali è spesso la sfida computazionale principale nell'inferenza bayesiana, specialmente quando lo spazio dei parametri ha alta dimensionalità.

Una delle proprietà più notevoli e utili dell'inferenza bayesiana è che possiamo incorporare nuove evidenze in modo sequenziale, aggiornando le nostre credenze passo dopo passo man mano che arrivano nuove informazioni. Questa proprietà rende l'apprendimento bayesiano naturalmente adatto per sistemi dinamici che acquisiscono dati nel tempo.

Supponiamo di osservare prima l'evidenza B_1 e di aggiornare le nostre credenze sull'ipotesi A usando il teorema di Bayes:

$$\mathbf{P(A|B_1) = [P(B_1|A) \cdot P(A)] / P(B_1)}$$

Questa è la nostra distribuzione posteriore dopo la prima osservazione. Ora supponiamo di osservare una seconda evidenza B_2. Come dovremmo aggiornare ulteriormente le nostre credenze? Applichiamo di nuovo il teorema di Bayes, ma ora usando $P(A|B_1)$ come nostro nuovo prior:

$$\mathbf{P(A|B_1, B_2) = [P(B_2|A, B_1) \cdot P(A|B_1)] / P(B_2|B_1)}$$

Il posteriore del passo precedente diventa il prior del passo corrente. Possiamo espandere completamente questa espressione sostituendo l'espressione per $P(A|B_1)$:

$$P(A|B_1, B_2) = [P(B_2|A, B_1) \cdot P(B_1|A) \cdot P(A)] / [P(B_1) \cdot P(B_2|B_1)]$$

Se B_1 e B_2 sono condizionalmente indipendenti dato A (cioè $P(B_2|A, B_1) = P(B_2|A)$), questa si semplifica ulteriormente. Ma anche senza questa assunzione, la struttura è chiara: ogni nuova evidenza modifica il posteriore moltiplicandolo per il fattore di verosimiglianza della nuova evidenza e rinormalizzando.

Proprietà fondamentale: L'ordine in cui riceviamo le evidenze non importa per il risultato finale (assumendo che tutte le evidenze siano eventualmente incorporate). Se osserviamo prima B_1 e poi B_2, o prima B_2 e poi B_1, otteniamo lo stesso posteriore finale $P(A|B_1, B_2)$.

Inoltre, non abbiamo bisogno di mantenere in memoria tutta la storia delle osservazioni. Il posteriore corrente è una **statistica sufficiente** per tutto ciò che abbiamo appreso finora. Questa è una forma della proprietà di Markov: il futuro dipende dal passato solo attraverso lo stato presente.

Verosimiglianza, Odds e il Rapporto di Verosimiglianza

Un modo particolarmente illuminante di comprendere l'aggiornamento bayesiano è attraverso il concetto di **odds** (rapporto di quote) e il **rapporto di verosimiglianza**.

Consideriamo due ipotesi alternative A_1 e A_2. L'odds a priori in favore di A_1 rispetto ad A_2 è:

Odds_prior(A_1:A_2) = P(A_1) / P(A_2)

Questo rapporto esprime quanto più probabile credevamo che fosse A_1 rispetto ad A_2 prima di osservare qualsiasi evidenza.

Similarmente, l'odds a posteriori dopo aver osservato l'evidenza B è:

Odds_post(A_1:A_2|B) = P(A_1|B) / P(A_2|B)

Applicando il teorema di Bayes a entrambi i termini:

$P(A_1|B) / P(A_2|B) = [P(B|A_1) \cdot P(A_1) / P(B)] / [P(B|A_2) \cdot P(A_2) / P(B)]$

I termini P(B) al denominatore si cancellano, lasciandoci con:

Odds_post = [P(B|A_1) / P(B|A_2)] × [P(A_1) / P(A_2)]

Riconosciamo il secondo fattore come gli odds a priori. Il primo fattore è chiamato **rapporto di verosimiglianza** (likelihood ratio):

LR(B) = P(B|A_1) / P(B|A_2)

Quindi possiamo scrivere elegantemente:

Odds_posteriori = Rapporto di Verosimiglianza × Odds_priori

Questa è la **forma multiplicativa del teorema di Bayes**. L'evidenza B aggiorna le nostre credenze relative moltiplicando gli odds priori per il rapporto di verosimiglianza:

- Se LR > 1: l'evidenza supporta A_1 rispetto ad A_2
- Se LR < 1: l'evidenza supporta A_2 rispetto ad A_1
- Se LR = 1: l'evidenza non discrimina tra le ipotesi

Il rapporto di verosimiglianza ha un'interpretazione naturale: misura quanto più fortemente l'evidenza B era prevista da A_1 rispetto ad A_2. È "oggettivo" nel senso che dipende solo dalle verosimiglianze, non dai prior.

Per osservazioni multiple indipendenti B_1, B_2, ..., B□, i rapporti di verosimiglianza si moltiplicano:

LR(B_1, ..., B□) = LR(B_1) × LR(B_2) × ... × LR(B□)

Scala logaritmica: Lavorare con logaritmi è spesso conveniente. Il logaritmo dell'odds è chiamato **logit**, e il logaritmo del rapporto di verosimiglianza è il **weight of evidence**:

log(Odds_post) = log(LR) + log(Odds_prior)

In questa forma additiva, ciascuna osservazione contribuisce additivamente al log-odds. Questo ha un'interpretazione informazionale: ciascuna osservazione aggiunge (o sottrae) una certa quantità di "evidenza" misurata in unità di informazione.

Il Fattore di Bayes e la Selezione dei Modelli

Un problema fondamentale nella scienza è il confronto tra modelli o teorie alternative. Il **fattore di Bayes** fornisce un framework bayesiano rigoroso per la selezione dei modelli.

Consideriamo due modelli M_1 e M_2. Ciascun modello specifica una verosimiglianza: $M_1 \rightarrow p(x|M_1)$ e $M_2 \rightarrow p(x|M_2)$. Se i modelli hanno parametri interni θ_1 e θ_2 con distribuzioni a priori $p(\theta_1|M_1)$ e $p(\theta_2|M_2)$, allora:

$$\mathbf{p(x|M_1) = \int p(x|\theta_1, M_1) \cdot p(\theta_1|M_1)\, d\theta_1}$$

$$\mathbf{p(x|M_2) = \int p(x|\theta_2, M_2) \cdot p(\theta_2|M_2)\, d\theta_2}$$

Il **fattore di Bayes** in favore di M_1 rispetto a M_2 è:

$$\mathbf{BF_{12} = p(x|M_1) / p(x|M_2)}$$

Il fattore di Bayes confronta quanto bene i due modelli predicono i dati, mediando sull'incertezza nei loro parametri.

Rasoio di Occam Bayesiano: Un modello più complesso con più parametri può adattarsi meglio ai dati, ma deve "spalmare" la sua massa probabilistica su uno spazio di parametri più grande, il che penalizza automaticamente la complessità. Questa è la **penalità di complessità automatica**.

Usando l'approssimazione di Laplace per grandi campioni:

$$\mathbf{\log BF_{12} \approx [\log p(x|\hat{\theta}_1, M_1) - \log p(x|\hat{\theta}_2, M_2)] - [(k_1 - k_2)/2] \log n}$$

dove:

- $\hat{\theta}_1$, $\hat{\theta}_2$ sono le stime di massima verosimiglianza
- k_1, k_2 sono i numeri di parametri nei due modelli
- n è la dimensione del campione

Il primo termine premia l'**adattamento** (fit) ai dati. Il secondo termine penalizza la **complessità** proporzionalmente a k log n.

Scala di interpretazione di Jeffreys per i fattori di Bayes:

BF_{12}	**Interpretazione**
< 1	Evidenza a favore di M_2
1 - 3	Evidenza debole per M_1
3 - 10	Evidenza sostanziale per M_1
10 - 30	Evidenza forte per M_1
30 - 100	Evidenza molto forte per M_1

> 100	Evidenza decisiva per M_1

Indipendenza Condizionale e Struttura Grafica

L'**indipendenza condizionale** è cruciale nella probabilità e nell'inferenza bayesiana. Due eventi A e B sono condizionalmente indipendenti dato C se:

P(A, B|C) = P(A|C) · P(B|C)

Equivalentemente: **P(A|B, C) = P(A|C)**

Sapere B non fornisce informazione addizionale su A una volta che conosciamo C.

L'indipendenza condizionale può essere rappresentata elegantemente usando **grafi**. In una **rete bayesiana**:

- I nodi rappresentano variabili aleatorie
- Gli archi diretti rappresentano dipendenze dirette
- L'assenza di archi implica indipendenza condizionale

La **fattorizzazione** della distribuzione congiunta segue la struttura del grafo. Se $X_1, ..., X_n$ sono le variabili e $Pa(X_i)$ denota i "genitori" di X_i nel grafo:

$$p(X_1, ..., X_n) = \prod_{i=1}^{n} p(X_i|Pa(X_i))$$

Questa fattorizzazione riduce drasticamente il numero di parametri. Una distribuzione congiunta di n variabili binarie richiederebbe $2^n - 1$ parametri senza struttura. Con una rete bayesiana appropriata, il numero di parametri può crescere solo linearmente con n.

Il Teorema di Rappresentazione di de Finetti

Uno dei risultati più profondi nella fondazione della probabilità bayesiana è il **teorema di rappresentazione di de Finetti**.

Una sequenza infinita di variabili aleatorie $X_1, X_2, X_3, ...$ si dice **scambiabile** se la distribuzione congiunta di qualsiasi sottogruppo finito è invariante rispetto a permutazioni:

$$p(X_1 = x_1, ..., X_n = x_n) = p(X_{\sigma(1)} = x_1, ..., X_{\sigma(n)} = x_n)$$

per qualsiasi permutazione σ.

Teorema di de Finetti: Se $X_1, X_2, X_3, ...$ sono scambiabili, allora esiste una variabile aleatoria θ e una distribuzione μ su θ tale che:

$$p(X_1 = x_1, ..., X_n = x_n) = \int \left[\prod_{i=1}^{n} p(X_i = x_i|\theta)\right] d\mu(\theta)$$

In parole: una sequenza scambiabile può sempre essere rappresentata come condizionalmente i.i.d. data una variabile latente θ, con una distribuzione a priori μ su θ.

Implicazione fondamentale: L'approccio bayesiano di introdurre parametri aleatori non è un artificio matematico arbitrario, ma una conseguenza necessaria dell'assunzione naturale di scambiabilità. Se crediamo che l'ordine delle osservazioni non dovrebbe importare, allora dobbiamo logicamente adottare un framework bayesiano.

La Questione Filosofica del A Priori

La scelta della distribuzione a priori è l'aspetto più discusso e controverso dell'inferenza bayesiana. Rappresenta il punto dove le assunzioni soggettive entrano esplicitamente nell'analisi.

Domande fondamentali:

- Da dove vengono i prior?
- Come dovremmo sceglierli?
- Esistono prior "corretti" o "oggettivi"?
- O sono inevitabilmente soggettivi?

Due scuole di pensiero:

1. Bayesiani Soggettivisti (de Finetti, Savage)

- Abbracciano la soggettività del prior
- La probabilità è sempre soggettiva – rappresenta il grado di credenza personale
- Non ha senso parlare di "la" probabilità in astratto
- La coerenza è garantita dagli assiomi della probabilità

2. Bayesiani Oggettivisti (Jeffreys, Jaynes)

- Cercano principi per costruire prior "oggettivi" o "non-informativi"
- Obiettivo: rappresentare uno stato di ignoranza razionale
- Eliminare l'arbitrarietà nella scelta del prior
- Rendere l'inferenza più "scientifica"

Prior Non-Informativi e il Problema della Parametrizzazione

Il concetto più ingenuo di prior non-informativo è il **prior uniforme**: se non sappiamo nulla su un parametro $\theta \in [a,b]$, usiamo $p(\theta) = \text{costante}$.

Problemi:

1. **Parametri illimitati:** Se $\theta \in (-\infty, \infty)$, una distribuzione uniforme sarebbe "impropria" (non integra a uno).

2. **Non-invarianza:** Il prior uniforme non è invariante rispetto a riparametrizzazioni. Se θ ha prior uniforme e $\varphi = g(\theta)$, il prior indotto su φ non è generalmente uniforme.

Usando la trasformazione di variabili:

$p(\varphi) = p(\theta) \cdot |d\theta/d\varphi|$

Se $p(\theta)$ è costante, $p(\varphi) \propto |dg^{-1}(\varphi)/d\varphi|$, che generalmente non è costante.

Paradosso di Bertrand: Questo illustra il problema. Per una corda casuale su un cerchio, ottengo risposte diverse (1/2, 1/3, o 1/4) a seconda di come interpreto "casualmente". La scelta della parametrizzazione non dovrebbe influenzare le conclusioni sostantive, ma il prior uniforme dà risposte diverse!

Il Prior di Jeffreys: Invarianza attraverso la Metrica di Fisher

Harold Jeffreys propose una soluzione elegante usando la **geometria intrinseca** dello spazio dei parametri determinata dall'**informazione di Fisher**.

L'informazione di Fisher per un parametro θ è:

$I(\theta) = -E[d^2 \log p(x|\theta)/d\theta^2] = E[(d \log p(x|\theta)/d\theta)^2]$

Misura quanto "informativa" è un'osservazione rispetto a θ.

Il **prior di Jeffreys** è:

$p(\theta) \propto \sqrt{I(\theta)}$

Per modelli multiparametro con vettore θ e matrice di informazione $I(\theta)$:

$p(\theta) \propto \sqrt{\det(I(\theta))}$

Proprietà cruciale: Il prior di Jeffreys è invariante sotto riparametrizzazioni. Se riparametrizziamo con $\varphi = g(\theta)$, il prior indotto su φ ha esattamente la stessa forma del prior di Jeffreys per φ.

Limitazioni:

- Per alcuni modelli risulta improprio
- Per modelli multidimensionali: "paradosso di marginizzazione"
- In alcuni casi porta a posteriori con proprietà indesiderabili

Prior di Riferimento e Maximum Entropy

Reference priors (Bernardo, Berger): Costruire prior che massimizzino l'informazione attesa che i dati forniranno sul parametro, misurata dalla divergenza di Kullback-Leibler tra prior e posteriore.

Principio di Massima Entropia (Jaynes): Data informazione sotto forma di vincoli, scegliere la distribuzione con massima entropia soggetta a quei vincoli.

L'entropia misura "dispersione" o "incertezza". Massimizzare l'entropia significa essere il meno specifici possibile compatibilmente con i vincoli noti.

Esempi:

- Supporto [a,b], nessun altro vincolo → Uniforme su [a,b]
- Media μ e varianza σ^2 note → Normale $N(\mu, \sigma^2)$
- Non-negativa con media μ → Esponenziale con rate 1/μ

Prior Coniugati: Eleganza Matematica

Una famiglia di prior si dice **coniugata** rispetto a una verosimiglianza se il posteriore appartiene alla stessa famiglia parametrica del prior.

Esempio classico: Beta-Binomiale

Verosimiglianza binomiale (k successi in n prove): $\mathbf{p(k|\theta, n) \propto \theta^k(1-\theta)^{n-k}}$

Prior Beta: $\mathbf{p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}}$

Posteriore Beta: $\mathbf{p(\theta|k, n) \propto \theta^{\alpha+k-1}(1-\theta)^{\beta+n-k-1}}$

I parametri posteriori sono: $\alpha_{post} = \alpha + k$, $\beta_{post} = \beta + n - k$

Altri esempi classici:

- **Gamma-Poisson:** Prior Gamma per rate λ di Poisson
- **Normale-Normale:** Prior Normale per media μ con varianza nota
- **Normal-Inverse-Gamma:** Prior congiunto per media e varianza
- **Dirichlet-Multinomial:** Prior Dirichlet per probabilità multinomiali

Vantaggi: Calcoli algebrici semplici, intuizione analitica

Limitazioni: La coniugazione vincola la forma funzionale del prior, che potrebbe non riflettere la vera conoscenza pregressa

Prior Informativi: Incorporare Conoscenza Sostanziale

In applicazioni reali, spesso abbiamo conoscenza pregressa:

- Risultati di studi precedenti
- Teorie scientifiche
- Esperienza pratica
- Opinioni di esperti

I **prior informativi** permettono di incorporare questa conoscenza formalmente.

Esempio: Stimare l'efficacia di un nuovo farmaco antiipertensivo. Sappiamo che farmaci simili tipicamente riducono la pressione di 10-20 mmHg con deviazione standard ~5 mmHg. Prior informativo: $N(15, 5^2)$.

Tecniche:

- **Meta-analisi bayesiana:** Usare studi precedenti per costruire il prior
- **Elicitazione del prior:** Interviste strutturate con esperti
- **Prior gerarchici:** I parametri del prior hanno a loro volta iperprior

Prior Debolmente Informativi e Regolarizzazione

Un compromesso pratico: **prior debolmente informativi** esprimono vincoli generali plausibili senza specificare troppo precisamente.

Esempi:

- Parametro positivo → Half-Cauchy o log-Normale con varianza grande
- Esclude valori impossibili ma è altrimenti vago

Connessione con Machine Learning:

- Prior gaussiani ↔ Regolarizzazione L2 (ridge regression)
- Prior Laplace ↔ Regolarizzazione L1 (lasso)
- Prior spike-and-slab → Sparsity

La prospettiva bayesiana fornisce un'interpretazione probabilistica coerente delle tecniche di regolarizzazione.

Natura e Interpretazione della Verosimiglianza

La funzione di verosimiglianza $\mathbf{L(\theta|x) = p(x|\theta)}$, vista come funzione di θ per dati fissi x, occupa un posto unico nell'inferenza statistica. A differenza di una distribuzione di probabilità, la verosimiglianza non è necessariamente normalizzata – l'integrale di L(θ|x) rispetto a θ generalmente non è uno.

Ronald Fisher sosteneva che la verosimiglianza contenesse tutta l'informazione rilevante nei dati per l'inferenza su θ.

Principio di verosimiglianza: Se due esperimenti producono verosimiglianze proporzionali, allora dovrebbero portare alle stesse conclusioni inferenziali su θ.

Nel framework bayesiano, la verosimiglianza entra direttamente nel teorema di Bayes come uno dei due ingredienti chiave (insieme al prior) che determinano il posteriore. La verosimiglianza rappresenta l'**influenza dei dati**, mentre il prior rappresenta la **conoscenza pregressa**. Il posteriore emerge come **sintesi bilanciata** di entrambi.

Costruzione della Verosimiglianza per Osservazioni Indipendenti

Per n osservazioni indipendenti x_1, x_2, ..., x_n, ciascuna con distribuzione f(x|θ), l'assunzione di indipendenza condizionale data θ permette la fattorizzazione:

$$\mathbf{L(\theta|x_1, ..., x_n) = \prod_{i=1}^{n} f(x_i|\theta)}$$

Per stabilità numerica e convenienza analitica, lavoriamo con la **log-verosimiglianza**:

$\ell(\theta|x) = \log L(\theta|x) = \Sigma_{i=1}^{n} \log f(x_i|\theta)$

Il logaritmo:

- Trasforma prodotti in somme
- Evita underflow numerico
- Preserva la stessa informazione (monotono crescente)

Risultato asintotico fondamentale: Sotto condizioni di regolarità, per grandi campioni la distribuzione della stima di massima verosimiglianza è approssimativamente normale:

$\hat{\theta}_{MLE} \sim N(\theta_{vero}, I(\theta_{vero})^{-1})$

dove $I(\theta)$ è l'informazione di Fisher.

Esempio Dettagliato: Modello Binomiale

Consideriamo n lanci di una moneta con probabilità θ di testa. Osserviamo k teste.

Verosimiglianza binomiale: $L(\theta|k, n) = (n \text{ choose } k)\ \theta^k(1-\theta)^{n-k}$

Il termine binomiale non dipende da θ, quindi: **$L(\theta|k, n) \propto \theta^k(1-\theta)^{n-k}$**

Log-verosimiglianza: $\ell(\theta|k, n) = k \log \theta + (n-k) \log(1-\theta) + \text{costante}$

Massima verosimiglianza: Deriviamo e poniamo uguale a zero:

$d\ell/d\theta = k/\theta - (n-k)/(1-\theta) = 0$

Risolvendo: **$k(1-\theta) = (n-k)\theta\ k = n\theta\ \hat{\theta} = k/n$**

La stima di massima verosimiglianza è la proporzione osservata di successi.

Verifica che sia un massimo: $d^2\ell/d\theta^2 = -k/\theta^2 - (n-k)/(1-\theta)^2 < 0$

La derivata seconda è negativa per tutti $\theta \in (0,1)$, confermando il massimo.

Informazione di Fisher: $I(\theta) = n/[\theta(1-\theta)]$

L'informazione:

- Cresce linearmente con n (più dati = più informazione)
- È massima quando θ = 0.5
- È minima agli estremi (θ vicino a 0 o 1)

Verosimiglianza Profile e Parziale

In problemi con molteplici parametri $\theta = (\psi, \lambda)$, dove ψ è di interesse e λ sono parametri di disturbo (nuisance parameters), tre approcci principali:

1. Verosimiglianza integrata (bayesiana): Specifichiamo un prior su λ e integriamo:

$L_{int}(\psi|x) = \int L(\psi, \lambda|x) \cdot p(\lambda)\, d\lambda$

Completamente bayesiano e coerente, ma richiede prior per λ.

2. Verosimiglianza profile: Per ciascun valore di ψ, ottimizziamo su λ:

$L_{prof}(\psi|x) = \max_\lambda L(\psi, \lambda|x) = L(\psi, \hat{\lambda}(\psi)|x)$

dove $\hat{\lambda}(\psi)$ è il valore di λ che massimizza la verosimiglianza per quel ψ fissato.

3. Verosimiglianza condizionale: Condizioniamo su una statistica sufficiente S per λ:

$L_{cond}(\psi|x, S) = L(\psi, \lambda|x, S)$

Elegante quando una statistica sufficiente esiste.

Verosimiglianza per Dati Censurati e Mancanti

Dati censurati (analisi di sopravvivenza):

Se osserviamo evento al tempo t_i:

- Contributo: densità $f(t_i|\theta)$

Se osservazione censurata al tempo c_i:

- Contributo: funzione di sopravvivenza $S(c_i|\theta) = P(T > c_i|\theta)$

Verosimiglianza totale: $L(\theta|dati) = [\prod_{osservati} f(t_i|\theta)] \times [\prod_{censurati} S(c_i|\theta)]$

Meccanismi di dati mancanti:

Missing Completely At Random (MCAR):

- P(mancante) non dipende né da valori osservati né mancanti
- Possiamo analizzare solo dati completi (complete-case analysis)

Missing At Random (MAR):

- P(mancante) dipende da dati osservati ma non da quelli mancanti
- L'analisi bayesiana può trattare dati mancanti come parametri latenti

Missing Not At Random (MNAR):

- P(mancante) dipende dai valori mancanti stessi
- Richiede modellare esplicitamente il meccanismo di missingness

Verosimiglianza sotto MAR: $L(\theta|x_{obs}) = \int L(\theta|x_{obs}, x_{miss}) \cdot p(x_{miss}|x_{obs}, \theta)\, dx_{miss}$

Modelli Gerarchici e Verosimiglianza Marginale

Modello a due livelli:

Livello 1 (osservazioni): $y_{ij} \sim N(\theta_j, \sigma^2)$ per i = 1,...,n_j, j = 1,...,J

Livello 2 (gruppi): $\theta_j \sim N(\mu, \tau^2)$ per j = 1,...,J

Verosimiglianza completa: $p(y|\theta_1, ..., \theta_J, \sigma^2) = \prod_j \prod_i (1/\sqrt{(2\pi\sigma^2)}) \exp[-(y_{ij} - \theta_j)^2/(2\sigma^2)]$

Verosimiglianza marginale: Marginalizzando sui parametri di gruppo θ_j:

$p(y|\mu, \tau^2, \sigma^2) = \prod_j \int [\prod_i p(y_{ij}|\theta_j, \sigma^2)] \cdot p(\theta_j|\mu, \tau^2)\, d\theta_j$

Per il caso normale-normale, la distribuzione marginale è: $y_{ij} \sim N(\mu, \sigma^2 + \tau^2)$

con correlazione intra-gruppo: $Corr(y_{ij}, y_{i'j}) = \tau^2/(\sigma^2 + \tau^2)$ per i ≠ i'

Shrinkage/Partial Pooling: Nei modelli gerarchici bayesiani, i parametri di gruppo "prendono in prestito forza" l'uno dall'altro attraverso il livello di popolazione. Gruppi con pochi dati sono tirati verso la media di popolazione, mentre gruppi con molti dati rimangono vicini alle loro medie osservate.

Robustezza della Verosimiglianza: Outliers e Misspecificazione

La verosimiglianza gaussiana standard è sensibile agli outliers perché la distribuzione normale ha code che decrescono esponenzialmente rapidamente.

Approcci per robustezza:

1. Verosimiglianze a code pesanti: Usare la **distribuzione t di Student** con v gradi di libertà:

$f(x|\mu, \sigma, v) \propto [1 + (x-\mu)^2/(v\sigma^2)]^{-(v+1)/2}$

Per piccoli v: code molto più pesanti della normale → meno peso agli outliers Al limite $v \to \infty$: converge alla normale

2. Modelli di mistura: $f(x|\theta) = (1-\varepsilon) f_{principale}(x|\theta) + \varepsilon f_{outlier}(x|\theta)$

dove ε è la proporzione di outliers. Il modello identifica automaticamente osservazioni anomale.

3. M-estimatori e verosimiglianze pesate: Modificare la verosimiglianza per ridurre l'influenza di osservazioni estreme (funzione di Huber, ecc.)

4. Analisi di sensibilità: Condurre multiple analisi variando le assunzioni e verificare la stabilità delle conclusioni.

Misspecificazione del modello:

Se il vero processo generatore dei dati non appartiene alla famiglia di modelli considerata, cosa stimano MLE o media posteriore?

Risposta: Sotto misspecificazione, questi stimatori convergono al valore del parametro che **minimizza la divergenza di Kullback-Leibler** tra il vero processo e il modello specificato. È la "miglior approssimazione" al processo vero all'interno della famiglia specificata.

Implicazione: Gli intervalli di confidenza/credibilità standard potrebbero sottostimare l'incertezza vera perché non tengono conto dell'incertezza del modello. Metodi come il "sandwich estimator" o model averaging bayesiano tentano di correggere.

Significato e Interpretazione del Posteriore

La distribuzione posteriore **p(θ|x)** rappresenta la sintesi completa di tutto ciò che possiamo sapere sul parametro θ dati i dati x e la nostra conoscenza pregressa.

Teorema di Bayes: $p(\theta|x) = [p(x|\theta) \cdot p(\theta)] / p(x)$

O in forma proporzionale: $p(\theta|x) \propto p(x|\theta) \cdot p(\theta)$

Posteriore ∝ Verosimiglianza × Prior

Interpretazioni:

- **Soggettivista:** Grado di credenza personale aggiornato dell'analista
- **Oggettivista:** Grado razionale di credenza che qualsiasi agente dovrebbe avere

Il posteriore permette calcoli esatti di:

- Probabilità che θ cada in qualsiasi regione
- Media, mediana, moda posteriore
- Intervalli di credibilità
- Qualsiasi altra statistica di interesse

Struttura del Posteriore: Compromesso tra A Priori e Verosimiglianza

Esempio illuminante: Normale con varianza nota

Supponiamo:

- $x \sim N(\theta, \sigma^2)$ dove σ^2 è noto
- $\theta \sim N(\mu_0, \tau_0^2)$ (prior)

Verosimiglianza: $p(x|\theta) \propto \exp[-(x-\theta)^2/(2\sigma^2)]$

Prior: $p(\theta) \propto \exp[-(\theta-\mu_0)^2/(2\tau_0^2)]$

Posteriore (anche normale): $\theta|x \sim N(\mu□, \tau□^2)$

dove:

$1/\tau□^2 = 1/\tau_0^2 + 1/\sigma^2$ (precisioni si sommano)

$\mu_n = (\mu_0/\tau_0^2 + x/\sigma^2) \cdot \tau_n^2$ (media pesata)

Riscrivendo la media posteriore:

$\mu_n = w_0\mu_0 + w_1 x$

dove:

- $w_0 = \sigma^2/(\tau_0^2 + \sigma^2)$
- $w_1 = \tau_0^2/(\tau_0^2 + \sigma^2)$
- $w_0 + w_1 = 1$

Interpretazione: La media posteriore è una media pesata tra la media prior μ_0 e l'osservazione x. I pesi dipendono dalle varianze relative:

- Prior preciso (τ_0^2 piccolo) → w_0 grande → posteriore vicino al prior
- Dati precisi (σ^2 piccolo) → w_1 grande → posteriore dominato dai dati

Per n osservazioni indipendenti:

$1/\tau_n^2 = 1/\tau_0^2 + n/\sigma^2$

$\mu_n = (\mu_0/\tau_0^2 + n\bar{x}/\sigma^2) \cdot \tau_n^2$

Quando $n \to \infty$:

- Il termine n/σ^2 domina
- $\mu_n \to \bar{x}$ (media campionaria)
- $\tau_n^2 \to \sigma^2/n$
- L'influenza del prior sparisce ("washing out")

Principio generale: Con dati sufficienti, prior ragionevoli diversi convergono a posteriori simili. Le differenze nel prior diventano irrilevanti con abbondanza di dati informativi.

Statistiche Sommarie del Posteriore

1. Media Posteriore: $\bar{\theta} = E[\theta|x] = \int \theta\, p(\theta|x)\, d\theta$

Stima di Bayes sotto perdita quadratica. Minimizza $E[(\theta-\delta)^2|x]$.

2. Mediana Posteriore: θ_med tale che **$P(\theta < \theta_med|x) = 0.5$**

Stima di Bayes sotto perdita assoluta. Minimizza $E[|\theta-\delta||x]$.

3. Moda Posteriore (MAP - Maximum A Posteriori): $\theta_MAP = \text{argmax}\ p(\theta|x)$

Stima di Bayes sotto perdita 0-1. Valore singolo più probabile.

4. Varianza Posteriore: $Var[\theta|x] = E[(\theta - E[\theta|x])^2|x]$

Misura la dispersione/incertezza residua.

5. Deviazione Standard Posteriore: SD[θ|x] = √Var[θ|x]

Incertezza nelle unità originali del parametro.

Intervalli di Credibilità e Regioni HPD

Un **intervallo di credibilità** al (1-α)×100% è [θ_L, θ_U] tale che:

P(θ_L < θ < θ_U | x) = 1 - α

Interpretazione diretta: C'è probabilità (1-α) che il vero parametro cada nell'intervallo, dati i dati osservati.

Contrasto con intervallo di confidenza frequentista: Se ripetessimo l'esperimento molte volte, circa (1-α)×100% degli intervalli conterrebbero il parametro. Per un particolare intervallo osservato, non possiamo dire che P(parametro nell'intervallo) = 1-α.

Intervallo equi-code (equal-tailed): P(θ < θ_L|x) = α/2 P(θ > θ_U|x) = α/2

Simmetrico e facile da calcolare usando quantili, ma non necessariamente ottimale.

Intervallo HPD (Highest Posterior Density): Il più corto intervallo con probabilità posteriore (1-α).

Formalmente, [θ_L, θ_U] che soddisfa:

1. **P(θ_L < θ < θ_U|x) = 1 - α**
2. **p(θ_L|x) = p(θ_U|x) = k** per qualche k
3. **p(θ|x) ≥ k** per tutti θ ∈ [θ_L, θ_U]

Tutti i punti dentro l'HPD hanno densità posteriore maggiore o uguale a qualsiasi punto fuori.

Proprietà:

- Per posteriori unimodali e simmetrici: HPD = equi-code
- Per posteriori asimmetrici: HPD è più corto (più informativo)
- Per posteriori multimodali: HPD potrebbe essere un'unione di intervalli disgiunti

Inferenza su Funzioni del Parametro

Se abbiamo p(θ|x) e siamo interessati a φ = g(θ), deriviamo p(φ|x) usando trasformazione di variabili.

Trasformazione univariata monotona: Se θ ha densità p(θ|x) e φ = g(θ) con inversa θ = $g^{-1}(φ)$:

$p(φ|x) = p(g^{-1}(φ)|x) \cdot |dg^{-1}(φ)/dφ|$

In pratica: Se abbiamo campioni dal posteriore di θ, semplicemente applichiamo g a ciascun campione per ottenere campioni dal posteriore di φ.

Esempio: Stimato log(θ), voglio inferenza su θ. Esponenzio ciascun campione di log(θ). La distribuzione empirica dei valori esponenziati approssima p(θ|x).

Potente per quantità derivate complesse: Regressione logistica con parametri α, β. Interessati a odds ratio exp(β). Trasformo campioni posteriori di β con exponenziale.

Predizione Posteriore

Distribuzione predittiva posteriore per nuova osservazione y_new dato x:

$p(y_new|x) = \int p(y_new|\theta)\, p(\theta|x)\, d\theta$

Mediamo la distribuzione predittiva p(y_new|θ) rispetto al posteriore su θ. Incorpora automaticamente l'incertezza sul parametro.

Contrasto con approccio "plug-in" frequentista: Stimare θ con $\hat{\theta}$ e usare $p(y_new|\hat{\theta})$. Ignora l'incertezza su θ, sottostimando l'incertezza predittiva.

Esempio normale: Dopo aver osservato x con posteriore $N(\mu_{\square}, \tau_{\square}^2)$, la predittiva per y_new è:

$y_new|x \sim N(\mu_{\square}, \sigma^2 + \tau_{\square}^2)$

Varianza predittiva = varianza dei dati (σ^2) + varianza posteriore del parametro ($\tau_{\square}^2$).

Usi della distribuzione predittiva:

1. Predizioni puntuali: E[y_new|x]
2. Intervalli predittivi
3. Model checking: confronto con osservazioni future
4. Validazione predittiva: calibrazione

Contrazioni del Posteriore e Consistenza

Cosa succede quando raccogliamo sempre più dati?

Sotto condizioni di regolarità, se il modello è corretto e il prior assegna probabilità positiva a intorni del vero θ_0:

$p(\theta|x_1, ..., x_{\square}) \rightarrow \delta(\theta_0)$ (massa puntuale su θ_0)

Il posteriore si "contrae" attorno al vero valore. La varianza posteriore tipicamente decresce come 1/n.

Teorema di Bernstein-von Mises: Sotto regolarità:

$\sqrt{n}(\theta - \theta_0)|x_1, ..., x_{\square} \rightarrow^d N(0, I(\theta_0)^{-1})$

Asintoticamente, il posteriore è approssimativamente normale, centrato sulla MLE, con varianza = inverso dell'informazione di Fisher.

Convergenza bayesiano-frequentista: In grandi campioni, la distribuzione posteriore bayesiana e la distribuzione campionaria frequentista della MLE diventano asintoticamente identiche.

Caveat: Vale solo asintoticamente e sotto modello corretto. In campioni finiti o con misspecificazione, le differenze possono essere sostanziali.

Posteriori Multimodali e Non-Normalità

In problemi complessi, il posteriore può essere multimodale o altamente non-normale.

Posteriore multimodale indica:

- I dati supportano multiple soluzioni distinte
- Può essere dovuto a simmetrie (es. label switching in modelli di mistura)
- Può riflettere genuina ambiguità nei dati

Esempio: Identificazione di sinusoide rumorosa. Se osserviamo solo parte di un ciclo, frequenze multiple potrebbero spiegare i dati quasi ugualmente bene → modi posteriori multipli.

Cautele necessarie:

- Statistiche sommarie come la media possono essere fuorvianti
- La media può cadere in regione di bassa probabilità tra modi
- Essenziale visualizzare l'intero posteriore
- Identificare tutti i modi principali e le loro probabilità relative

Sfide computazionali: Metodi MCMC standard possono rimanere intrappolati in un modo locale. Tecniche avanzate:

- Parallel tempering
- Simulated annealing
- Population MCMC

Posteriori Impropri

Quando usiamo **prior impropri** (che non integrano a uno), dobbiamo verificare che il posteriore risultante sia proprio.

Un prior improprio può portare a posteriore proprio se i dati forniscono sufficiente informazione.

Esempio: Prior improprio uniforme $p(\theta) \propto 1$ per $\theta \in (-\infty, \infty)$ combinato con verosimiglianza normale produce posteriore proprio con ≥1 osservazione.

Controesempio: Prior improprio $p(\sigma^2) \propto 1/\sigma^2$ e verosimiglianza normale con una sola osservazione → posteriore improprio per σ^2.

Regole generali per verificare proprietà:

1. Tail behavior: Il posteriore deve decadere sufficientemente rapidamente alle code
2. Dimensionalità: In alta dimensione, condizioni più stringenti
3. Verifica formale: Calcolare o limitare superiormente l'integrale normalizzante

Quando si usano prior impropri, è **essenziale verificare formalmente** la proprietà del posteriore prima di procedere con l'inferenza.

Fondamenti della Teoria della Decisione

La teoria della decisione bayesiana fornisce un framework formale per prendere decisioni ottimali in condizioni di incertezza. Estende l'inferenza bayesiana dall'obiettivo di apprendere sui parametri all'obiettivo di scegliere azioni che massimizzano l'utilità attesa o minimizzano la perdita attesa.

Un problema decisionale bayesiano è caratterizzato da quattro elementi:

1. **Spazio delle azioni A:** L'insieme di tutte le possibili azioni o decisioni disponibili
2. **Spazio dei parametri Θ:** L'insieme dei possibili stati del mondo o valori dei parametri
3. **Funzione di perdita L(θ, a):** Il costo di prendere l'azione a quando lo stato vero è θ
4. **Distribuzione posteriore p(θ|x):** La credenza aggiornata sui parametri dato i dati x

Obiettivo: Scegliere l'azione a* che minimizza la perdita attesa posteriore:

$a = argmin_{a \in A} \rho(a|x)^*$

dove la perdita attesa posteriore è:

$$\rho(a|x) = E[L(\theta, a)|x] = \int L(\theta, a)\, p(\theta|x)\, d\theta$$

L'azione ottimale a* è chiamata **decisione di Bayes** o **regola di Bayes**.

Questo framework è completamente generale e unifica molti problemi: stima puntuale, test di ipotesi, classificazione, predizione – tutti sono problemi decisionali con appropriate funzioni di perdita.

Funzioni di Perdita Comuni e le Loro Decisioni Ottimali

1. Perdita Quadratica: $L(\theta, \delta) = (\theta - \delta)^2$

Penalizza errori proporzionalmente al loro quadrato.

Decisione di Bayes: Media posteriore $\delta = E[\theta|x] = \int \theta\, p(\theta|x)\, d\theta^*$

Dimostrazione: Minimizziamo: $\rho(\delta|x) = E[(\theta - \delta)^2|x] = E[\theta^2|x] - 2\delta E[\theta|x] + \delta^2$

Derivando: $d\rho/d\delta = -2E[\theta|x] + 2\delta = 0$ $\delta = E[\theta|x]^*$

La derivata seconda è $2 > 0$, confermando il minimo.

2. Perdita Assoluta: $L(\theta, \delta) = |\theta - \delta|$

Penalizza errori linearmente. Più robusta agli outliers rispetto alla perdita quadratica.

Decisione di Bayes: Mediana posteriore *δ = mediana di p(θ|x)**

3. Perdita 0-1: L(θ, δ) = 0 se θ = δ, 1 altrimenti

Costo 1 se sbagliamo, 0 se indoviniamo. Appropriata per classificazione discreta.

Decisione di Bayes: Moda posteriore (MAP) *δ = argmax p(θ|x)**

4. Perdita LINEX: L(θ, δ) = exp[c(δ - θ)] - c(δ - θ) - 1

Perdita asimmetrica: sovrastime e sottostime hanno costi diversi, determinati dal segno e magnitudine di c.

Esempio: In ingegneria strutturale, sovrastimare la resistenza di un materiale (→ progettazione insufficiente) può essere catastrofico, mentre sottostimare (→ sovra-ingegnerizzazione) è semplicemente inefficiente.

5. Perdita Pesata: L(θ, δ) = w(θ)(θ - δ)²

Perdita quadratica dove il peso dipende dal vero stato θ. Permette di dare più importanza alla precisione in certe regioni dello spazio dei parametri.

Decisioni in Problemi di Classificazione

Nei problemi di classificazione, assegniamo un'osservazione x a una di K classi. Lo spazio delle azioni è discreto: **A = {1, 2, ..., K}**.

La **matrice di perdita** L(j, k) specifica il costo di classificare nella classe k quando la vera classe è j. Tipicamente:

- **L(j, j) = 0** (classificazione corretta)
- **L(j, k) > 0** per j ≠ k (errori)

Regola di Bayes generale: Classificare in classe k* dove:

*k = argmin_k Σ□ L(j, k) P(classe = j|x)**

Per perdita 0-1 (tutti gli errori hanno costo 1):

*k = argmax_k P(classe = k|x)**

Scegli la classe con massima probabilità posteriore (regola intuitiva).

Per matrici di perdita generali: La decisione ottimale bilancia probabilità posteriori con costi degli errori.

Esempio (diagnosi medica):

- **L(malato, diagnosi sano) >> L(sano, diagnosi malato)**

- Un falso negativo è molto più costoso di un falso positivo
- La regola di Bayes abbassa la soglia per diagnosticare la malattia

Analisi Costi-Benefici e Soglie Decisionali

Problema binario: Una condizione può essere presente (θ = 1) o assente (θ = 0). Decidiamo se dichiarare presente (a = 1) o assente (a = 0).

Costi:

- **C_FP:** costo di un falso positivo
- **C_FN:** costo di un falso negativo
- Costo zero per decisioni corrette

Perdita attesa per dichiarare presente: $\rho(a = 1|x) = C_{FP} \cdot P(\theta = 0|x)$

Perdita attesa per dichiarare assente: $\rho(a = 0|x) = C_{FN} \cdot P(\theta = 1|x)$

Dovremmo dichiarare presente se: $C_{FP} \cdot P(\theta = 0|x) < C_{FN} \cdot P(\theta = 1|x)$

Riorganizzando: **$P(\theta = 1|x) / P(\theta = 0|x) > C_{FP} / C_{FN}$**

Gli odds posteriori devono superare il rapporto dei costi.

Alternativamente: $P(\theta = 1|x) > C_{FP} / (C_{FP} + C_{FN})$

Questo stabilisce una **soglia di probabilità posteriore** per la decisione:

- Se C_FP = C_FN (costi uguali) → soglia = 0.5
- Se C_FN >> C_FP (falsi negativi molto più costosi) → soglia si abbassa

Questa analisi fornisce una base razionale e quantitativa per stabilire soglie decisionali. Ampiamente usata in:

- Medicina (screening diagnostici)
- Finanza (rilevamento frodi)
- Sicurezza (sistemi di allarme)

Valore dell'Informazione

EVPI (Expected Value of Perfect Information): Quantifica quanto saremmo disposti a pagare per eliminare completamente l'incertezza prima di decidere.

Decisione ora basata sul posteriore corrente: $\rho = \min_a E[L(\theta, a)|x]$*

Se potessimo apprendere il vero θ prima di decidere, per ogni θ sceglieremmo: $a(\theta) = \text{argmin}_a L(\theta, a)$*

Perdita attesa sotto informazione perfetta: $\rho_{perfetta} = E[\min_a L(\theta, a)|x] = E[L(\theta, a(\theta))|x]$*

Valore dell'informazione perfetta: *EVPI = ρ - ρ_perfetta**

Riduzione nella perdita attesa ottenibile eliminando tutta l'incertezza. Rappresenta il massimo da pagare per informazione perfetta.

EVSI (Expected Value of Sample Information): Valore di esperimenti aggiuntivi che riducono ma non eliminano l'incertezza.

Applicazioni:

- Progettazione ottimale di esperimenti
- Decisioni su quali dati raccogliere
- Se EVPI è piccolo → ulteriori dati probabilmente non cambieranno la decisione
- Se EVPI è grande → siamo in regione di alta incertezza, più informazione potrebbe essere preziosa

Regret e Minimax

Regret (rimpianto) di un'azione a nello stato θ:

R(θ, a) = L(θ, a) - min_{a'} L(θ, a')

Perdita addizionale rispetto alla decisione ottimale se conoscessimo θ. Sempre non-negativo, zero per l'azione ottimale dato θ.

Regret massimo: max_θ R(θ, a)

Strategia minimax regret: a_minimax = argmin_a max_θ R(θ, a)

Minimizza il rimpianto nel caso peggiore. Strategia conservativa appropriata quando:

- Non abbiamo informazione per costruire un prior significativo
- Vogliamo evitare decisioni che potrebbero risultare molto cattive in certi scenari

Connessioni profonde con teoria dei giochi e ottimizzazione robusta.

Analisi di Utilità Multi-Attributo

In problemi reali, le conseguenze hanno molteplici obiettivi conflittuali:

- Costo, rischio, efficacia
- Impatto ambientale
- Considerazioni etiche

Funzione di utilità multi-attributo: $U(x_1, x_2, ..., x_m)$

Assegna un valore di utilità complessivo a un vettore di m attributi.

Sotto condizioni di indipendenza, decomposizione: $U(x_1, ..., x_m) = \Sigma_i w_i u_i(x_i)$

dove:

- u_i: funzioni di utilità univariate per ciascun attributo
- w_i: pesi che riflettono l'importanza relativa

Decisione ottimale: Massimizza l'utilità attesa posteriore rispetto a tutti gli attributi.

Metodi di elicitazione:

- Analytic Hierarchy Process (AHP)
- Multi-Attribute Utility Theory (MAUT)
- Tecniche di elicitazione interattiva

La Sfida dell'Integrazione Multidimensionale

Il cuore del problema computazionale nell'inferenza bayesiana è il calcolo dell'**evidenza marginale** (costante di normalizzazione):

$$p(x) = \int p(x|\theta)\, p(\theta)\, d\theta$$

Questo integrale appare al denominatore del teorema di Bayes ed è essenziale per normalizzare il posteriore. Per modelli anche moderatamente complessi, questo integrale è analiticamente intrattabile.

La maledizione della dimensionalità: Se θ è k-dimensionale, l'integrale è k-dimensionale. Per k grande (k > 10), metodi numerici deterministici standard diventano impossibilmente costosi.

Manifestazioni:

- Lo spazio cresce esponenzialmente: n punti per dimensione → n^k punti totali
- La maggior parte dello spazio ha densità trascurabile
- In alta dimensione, quasi tutta la massa si concentra in un guscio sottile

Soluzione pratica: Per molte applicazioni non abbiamo bisogno di calcolare esattamente p(x). Possiamo lavorare con il posteriore non-normalizzato:

$$p(\theta|x) \propto p(x|\theta)\, p(\theta)$$

e usare metodi che campionano da questa distribuzione senza conoscere la costante di normalizzazione.

Approccio Monte Carlo: Campionamento come Soluzione

Idea fondamentale: Approssimare aspettazioni (integrali) usando medie campionarie.

Vogliamo calcolare: $\mathbf{E[g(\theta)|x]} = \int g(\theta)\, p(\theta|x)\, d\theta$

Con N campioni $\theta^{(1)}, \theta^{(2)}, ..., \theta^{(N)}$ dalla distribuzione posteriore:

$E[g(\theta)|x] \approx (1/N) \Sigma_{i=1}^{N} g(\theta^{(i)})$

Legge dei grandi numeri: Questa approssimazione converge quasi certamente al valore vero quando $N \to \infty$.

Teorema del limite centrale: L'errore decresce come $1/\sqrt{N}$:

$\sqrt{N}$ [approssimazione - valore vero] $\to^d N(0, \sigma^2)$

dove σ^2 è la varianza di $g(\theta)$ sotto $p(\theta|x)$.

Vantaggio cruciale: Il tasso di convergenza **$1/\sqrt{N}$ non dipende dalla dimensionalità k**. Enorme vantaggio rispetto ai metodi deterministici (es. quadratura) il cui costo cresce esponenzialmente con k.

Problema: Per la maggior parte dei posteriori interessanti, il campionamento diretto non è possibile. Soluzione: **MCMC**.

Markov Chain Monte Carlo: Principi Fondamentali

Idea geniale: Costruire una catena di Markov la cui distribuzione stazionaria (limite) sia esattamente il posteriore che vogliamo campionare.

Catena di Markov: Sequenza di stati aleatori dove ogni stato dipende solo dallo stato precedente. Definita da un **kernel di transizione** $K(\theta'|\theta)$.

Partendo da $\theta^{(0)}$, generiamo: **$\theta^{(1)} \sim K(\cdot|\theta^{(0)})$ $\theta^{(2)} \sim K(\cdot|\theta^{(1)})$ $\theta^{(3)} \sim K(\cdot|\theta^{(2)})$** ...

Sotto condizioni appropriate (irriducibilità, aperiodicità, esistenza di distribuzione stazionaria), la distribuzione di $\theta^{(t)}$ converge a una distribuzione limite $\pi(\theta)$ che non dipende dallo stato iniziale.

Obiettivo MCMC: Progettare K in modo che π sia esattamente il nostro posteriore $p(\theta|x)$.

Dopo un periodo di **"burn-in"** o **"warm-up"** durante il quale la catena converge, i campioni successivi possono essere trattati come campioni (approssimativamente) dalla distribuzione posteriore.

Algoritmo di Metropolis-Hastings

Sviluppato da Metropolis et al. (1953) per fisica statistica e generalizzato da Hastings (1970). È il metodo MCMC più generale e fondamentale.

Algoritmo:

1. **Inizializza** $\theta^{(0)}$ a qualche valore iniziale
2. **Per t = 0, 1, 2, ..., T:**
    a. **Proponi** θ^* campionando dalla distribuzione di proposta $q(\cdot|\theta^{(t)})$

b. **Calcola il rapporto di accettazione:**

$\alpha = min\{1, [p(\theta|x)\ q(\theta^{(t)}|\theta)] / [p(\theta^{(t)}|x)\ q(\theta^*|\theta^{(t)})]\}$**

c. **Campiona** u ~ Uniforme(0, 1)

d. **Se u < α:** accetta e poni $\theta^{(t+1)} = \theta^*$ **Altrimenti:** rigetta e poni $\theta^{(t+1)} = \theta^{(t)}$

Punto chiave: Il rapporto di accettazione può essere calcolato senza conoscere p(x), perché:

$p(\theta|x) / p(\theta^{(t)}|x) = [p(x|\theta)\ p(\theta^*)] / [p(x|\theta^{(t)})\ p(\theta^{(t)})]$**

I termini p(x) si cancellano.

Scelte comuni per q:

- **Random walk:** $q(\theta^*|\theta^{(t)}) = N(\theta^{(t)}, \Sigma)$
- **Indipendente:** $q(\theta^*|\theta^{(t)}) = q(\theta^*)$ (non dipende dallo stato corrente)
- **Adaptive:** Σ viene aggiustata durante il warm-up per ottimizzare l'accettazione

La scelta di q influenza l'**efficienza** ma non la **correttezza** – qualsiasi q soddisfacente condizioni minime convergerà al posteriore corretto.

Gibbs Sampling

Caso speciale di Metropolis-Hastings dove aggiorniamo una componente del vettore di parametri alla volta, campionando dalla sua **distribuzione condizionale completa** (full conditional) date tutte le altre componenti.

Per $\theta = (\theta_1, \theta_2, ..., \theta_k)$, un'iterazione del Gibbs sampler:

1. Campiona $\boldsymbol{\theta_1^{(t+1)} \sim p(\theta_1 \mid \theta_2^{(t)}, \theta_3^{(t)}, ..., \theta_k^{(t)}, x)}$
2. Campiona $\boldsymbol{\theta_2^{(t+1)} \sim p(\theta_2 \mid \theta_1^{(t+1)}, \theta_3^{(t)}, ..., \theta_k^{(t)}, x)}$
3. Campiona $\boldsymbol{\theta_3^{(t+1)} \sim p(\theta_3 \mid \theta_1^{(t+1)}, \theta_2^{(t+1)}, \theta_4^{(t)}, ..., \theta_k^{(t)}, x)}$

...

k. Campiona $\boldsymbol{\theta_k^{(t+1)} \sim p(\theta_k \mid \theta_1^{(t+1)}, \theta_2^{(t+1)}, ..., \theta_\{k-1\}^{(t+1)}, x)}$

Vantaggi:

- Non richiede tuning (nessun rapporto di accettazione)
- Tasso di accettazione = 1
- Semplice da implementare quando le full conditionals sono note

Software BUGS: Rese Gibbs sampling accessibile a non specialisti. Specifica il modello in linguaggio semplice, BUGS automaticamente costruisce e fa girare la catena di Markov appropriata.

Hamiltonian Monte Carlo (HMC)

Usa idee dalla fisica hamiltoniana per proporre mosse più efficienti nello spazio dei parametri. Particolarmente efficace per modelli ad alta dimensionalità.

Idea fisica: Trattare il negativo del log-posteriore come energia potenziale. Introdurre variabili momento ausiliarie. Simulare la dinamica hamiltoniana per proporre nuovi stati.

Vantaggio: Le proposte HMC possono fare grandi passi nello spazio dei parametri seguendo la geometria del posteriore, con alti tassi di accettazione.

No-U-Turn Sampler (NUTS): Estensione dell'HMC che adatta automaticamente i suoi parametri. Implementato nel software **Stan**, diventato lo standard de facto per inferenza bayesiana seria.

Diagnosi di Convergenza MCMC

Problemi potenziali:

- La catena non ha ancora raggiunto la distribuzione stazionaria
- Mixing povero (esplorazione lenta dello spazio)
- Correlazione tra campioni successivi

Strumenti diagnostici:

1. Trace plots: Visualizzare $\theta^{(t)}$ vs t. Dovrebbe mostrare "caterpillar" stabile, non trend o pattern.

2. R-hat statistic (Gelman-Rubin): Confronta varianza within-chain e between-chain. R-hat ≈ 1.00 indica convergenza. R-hat > 1.1 suggerisce problemi.

3. Effective Sample Size (ESS): Stima il numero di campioni indipendenti equivalenti. ESS molto minore di N indica alta autocorrelazione.

4. Autocorrelazione: Grafico di autocorrelazione dei campioni. Dovrebbe decadere rapidamente.

5. Multiple chains: Correre più catene con inizializzazioni diverse. Dovrebbero convergere alla stessa distribuzione.

Accelerazione e Ottimizzazione MCMC

Tecniche per migliorare l'efficienza:

1. Reparametrizzazione:

- Trasformare parametri per ridurre correlazioni posteriori
- Centramento vs non-centramento nei modelli gerarchici

2. Adaptive tuning:

- Durante warm-up, adattare la varianza della proposta
- Target: tasso di accettazione 20-30% (Metropolis) o 60-70% (HMC)

3. Parallel tempering:

- Correre più catene a diverse "temperature"
- Scambiare stati tra catene
- Aiuta a superare modi locali in posteriori multimodali

4. Thinning:

- Salvare solo ogni k-esimo campione per ridurre autocorrelazione
- Nota: generalmente meglio non thin e usare tutti i campioni

5. Rao-Blackwellization:

- Quando possibile, integrare analiticamente sottoinsiemi di parametri
- Riduce la varianza delle stime

Metodi di Approssimazione Variazionale

Alternativa a MCMC quando il campionamento è troppo lento. **Variational Inference (VI)** approssima il posteriore con una distribuzione più semplice dalla famiglia Q.

Idea: Trovare $q(\theta) \in Q$ che minimizza la divergenza KL dal vero posteriore:

$$q(\theta) = \text{argmin}_{q \in Q} KL(q(\theta) \,||\, p(\theta|x))$$

Equivalentemente, massimizza il **lower bound** sull'evidenza (ELBO):

$$\textbf{ELBO}(q) = E_q[\log p(x,\theta)] - E_q[\log q(\theta)]$$

Vantaggi:

- Spesso molto più veloce di MCMC
- Scalabile a dataset enormi
- Convergenza garantita (a un ottimo locale)

Svantaggi:

- Approssimazione potenzialmente povera
- Sottostima l'incertezza
- Nessuna garanzia di convergenza al vero posteriore

Varianti:

- **Mean-field VI:** Assume fattorizzazione completa $q(\theta) = \prod_i q_i(\theta_i)$
- **Automatic Differentiation VI (ADVI):** Usa gradients automatici
- **Normalizing Flows:** Trasformazioni più flessibili per q

Sfide Computazionali Future

Problemi aperti e direzioni di ricerca:

1. Scalabilità a Big Data:

- Dataset con milioni/miliardi di osservazioni
- Approximate methods: subsetting, variational inference
- Distributed/parallel MCMC

2. Modelli ad altissima dimensionalità:

- Spazi di parametri con migliaia/milioni di dimensioni
- Sparsity, regolarizzazione
- Metodi basati su gradienti (HMC, VI)

3. Modelli con likelihood intrattabili:

- Simulatori complessi senza formula chiusa
- Approximate Bayesian Computation (ABC)
- Likelihood-free inference

4. Quantificazione dell'incertezza in deep learning:

- Bayesian neural networks
- Dropout as approximation
- Laplace approximation per NN

5. Inferenza causale:

- Identificazione di relazioni causali
- Counterfactual reasoning
- Do-calculus bayesiano

La Macchina Enigma e Bletchley Park

Una delle applicazioni più drammatiche del ragionamento bayesiano avvenne durante la Seconda Guerra Mondiale. Alan Turing e colleghi a Bletchley Park svilupparono metodi per decifrare Enigma che incarnavano perfettamente i principi bayesiani, sebbene non esplicitamente formulati così all'epoca.

Il metodo "Banburismus" di Turing funzionava in questo modo ì: iniziavano con prior uniformi su ~159 quintilioni di configurazioni possibili della macchina. Per ogni messaggio intercettato, cercavano pattern statistici. Ogni pattern forniva evidenza che aggiornava le probabilità relative delle chiavi.

Il concetto di "ban": Turing introdusse il "ban" come unità di evidenza - essenzialmente logaritmi in base 10 del rapporto di verosimiglianza. Un ban significava che l'odds in favore di un'ipotesi era aumentato di un fattore 10:

Evidenza in bans = $\log_{10}(LR) = \log_{10}[P(dati|chiave_1)/P(dati|chiave_2)]$

Quando l'evidenza cumulativa raggiungeva 3-4 bans (odds di 1000:1 o 10000:1), procedevano con quella chiave. Questo aggiornamento bayesiano sequenziale permise agli Alleati di decifrare messaggi tedeschi regolarmente.

Impatto storico: Gli storici stimano che questo accorciò la guerra di 2-4 anni e salvò milioni di vite. Il ragionamento bayesiano, così astratto, ebbe impatto monumentale sulla storia mondiale.

Il Processo O.J. Simpson: La Fallacia del Procuratore

Il processo del 1995 fornì un esempio pubblico di ragionamento probabilistico scorretto. L'accusa presentò un'evidenza di abuso pregresso. La difesa controbatté: solo 1 su 2500 donne abusate viene poi uccisa dal partner, quindi l'evidenza era quasi irrilevante.

L'errore: Stavano calcolando P(omicidio|abuso) invece di P(colpevole|omicidio ∩ abuso).

Il ragionamento corretto: Dato che Nicole Brown Simpson è stata uccisa, la domanda rilevante è: tra le donne abusate che vengono poi uccise, quale frazione viene uccisa dal partner abusante? Questa probabilità è ~90%.

Formalmente, se H = "colpevole", E_1 = "omicidio", E_2 = "abuso":

$P(H|E_1,E_2) = P(E_1,E_2|H)P(H) / P(E_1,E_2)$

La difesa confondeva $P(E_1|E_2)$ con $P(H|E_1,E_2)$ - un errore fatale che il teorema di Bayes rende cristallino.

USS Scorpion: Ricerca Bayesiana negli Oceani

Nel 1968, il sottomarino nucleare USS Scorpion scomparve nell'Atlantico. John Craven applicò metodi bayesiani per guidare la ricerca in migliaia di miglia quadrate di oceano profondo.

Il metodo:

1. Generò scenari multipli (esplosione, collisione, guasto)
2. Per ogni scenario, calcolò dove il sottomarino sarebbe affondato
3. Assegnò prior a ciascuno scenario
4. Creò una mappa di probabilità dell'oceano

Man mano che aree venivano cercate senza successo, Craven aggiornava bayesianamente:

$P(posizione|non\ trovato\ in\ area\ A) \propto P(non\ trovato\ in\ A|posizione) \times P(posizione)$

Dove P(non trovato in A|posizione) = 0 se posizione ∈ A, altrimenti = 1.

Risultato: Lo Scorpion fu trovato a soli 200 metri dalla posizione di massima probabilità posteriore, in un'area inizialmente a bassa probabilità che era stata progressivamente rivalutata.

Questo successo portò all'adozione diffusa di metodi bayesiani in ricerca e salvataggio, successivamente usati per Air France 447 (2011).

Screening HIV: L'Importanza dei Prior in Medicina

Problema: Test HIV con sensibilità 99.9% e specificità 99.9%. Paziente a basso rischio risulta positivo. Qual è P(HIV|test+)?

Intuizione ingenua: "Test accurato al 99.9%, quindi ~99.9%"

Calcolo corretto: Con prevalenza P(HIV) = 0.3%:

P(HIV|+) = P(+|HIV)×P(HIV) / [P(+|HIV)×P(HIV) + P(+|¬HIV)×P(¬HIV)] = 0.999×0.003 / [0.999×0.003 + 0.001×0.997] = 0.002997 / 0.003994 = 0.75 = 75%

Solo 75%, non 99.9%! Il prior basso domina.

Implicazioni:

- Test positivi in screening richiedono sempre conferma
- Il valore predittivo dipende dalla prevalenza
- In popolazioni ad alto rischio (es. P(HIV)=10%), lo stesso test+ → P(HIV|+) ≈ 99.1%

Questo esempio mostra perché la comprensione bayesiana è fondamentale in medicina.

Spam Filtering: Bayesiano nel Quotidiano

Negli anni '90, Paul Graham pionierizzò filtri bayesiani per spam. Per ogni parola w:

- P(w|spam): probabilità in spam
- P(w|ham): probabilità in email legittime

Per nuova email con parole $w_1,...,w_n$:

$P(spam|w_1,...,w_n) \propto P(spam) \times \prod_i P(w_i|spam)$

Assumendo indipendenza condizionale (assunzione "naive" ma efficace).

Parole discriminanti:

- "Viagra", "enlargement" → alto P(w|spam)/P(w|ham)
- "meeting", "schedule" → basso ratio

Il classificatore raggiunse accuratezze >99% e divenne la base per Gmail, SpamAssassin e innumerevoli sistemi anti spami. Principi matematici del XVIII secolo risolvono problemi del XXI.

Entropia e Contenuto Informativo

Il legame tra teoria bayesiana e teoria dell'informazione è profondo. L'entropia di Shannon misura incertezza:

$H[p] = -\sum_i p(x_i) \log_2 p(x_i)$ (discreta) $H[p] = -\int p(x) \log p(x) \, dx$ (continua)

Interpretazione: Numero medio di bit per codificare un campione con codice ottimale.

Nel contesto bayesiano: L'entropia posteriore $H[p(\theta|x)]$ misura incertezza residua dopo aver osservato dati. L'aggiornamento bayesiano riduce entropia:

Informazione guadagnata = $H[p(\theta)] - H[p(\theta|x)] \geq 0$

I dati non possono aumentare incertezza in un framework coerente.

Divergenza di Kullback-Leibler

La divergenza KL misura quanto q differisce da p:

$KL(q||p) = \int q(x) \log[q(x)/p(x)] \, dx$

Proprietà:

- Sempre non-negativa: $KL(q||p) \geq 0$
- Zero $\Leftrightarrow$ q = p quasi ovunque
- Non simmetrica: $KL(q||p) \neq KL(p||q)$

Ruoli multipli in inferenza bayesiana:

1. **Variational Inference**: Approssima posteriore $p(\theta|x)$ con $q(\theta)$ minimizzando $KL(q||p)$

2. **Information Gain**: Informazione guadagnata da osservare x: $I(\theta;x) = KL(p(\theta|x)||p(\theta)) = H[\theta] - H[\theta|x]$

3. **Model Selection**: AIC approssima divergenza KL attesa tra modello stimato e vero processo

4. **Predictive Accuracy**: Log-likelihood predittiva legata a KL tra predittiva e distribuzione vera

Informazione di Fisher e Geometria Intrinseca

L'informazione di Fisher ha interpretazione geometrica profonda. Per parametro scalare:

$I(\theta) = E[(\partial \log p(x|\theta)/\partial\theta)^2] = -E[\partial^2 \log p(x|\theta)/\partial\theta^2]$

Misura quanto rapidamente log-likelihood cambia. Alta $I(\theta)$ → parametro ben identificabile.

Forma matriciale (parametri multipli):

$[I(\theta)]_{ij} = E[(\partial \log p(x|\theta)/\partial\theta_i)(\partial \log p(x|\theta)/\partial\theta_j)]$

Metrica di Fisher-Rao: Definisce distanza nello spazio parametri:

$ds^2 = \sum_{ij} I_{ij}(\theta)\, d\theta_i\, d\theta_j$

Questa metrica è invariante sotto riparametrizzazioni - cattura geometria intrinseca dello spazio statistico.

Il prior di Jeffreys $p(\theta) \propto \sqrt{\det(I(\theta))}$ è esattamente la misura di volume naturale per questa geometria - la densità "uniforme" rispetto alla metrica intrinseca.

Bound di Cramér-Rao

L'informazione di Fisher pone limite fondamentale sulla precisione di stima:

$\mathrm{Var}(\hat{\theta}) \geq 1/[n \cdot I(\theta)]$

per qualsiasi stimatore non distorto $\hat{\theta}$ basato su n osservazioni indipendenti.

Stimatori che raggiungono questo bound sono "efficienti". La MLE è asintoticamente efficiente.

Interpretazione bayesiana: Per prior non-informativi, la varianza posteriore approssima $1/[n \cdot I(\theta)]$, riflettendo che l'informazione viene principalmente dai dati.

Principio di Massima Entropia (Jaynes)

Principio: Data informazione parziale (vincoli), scegli la distribuzione con massima entropia. Razionale: sii il meno specifico possibile - non assumere info che non hai.

Formulazione: Massimizza $H[p]$ soggetto a:

- Normalizzazione: $\int p(x)dx = 1$
- Vincoli: $\int g_i(x)p(x)dx = \mu_i$

Soluzione (via Lagrange): Forma esponenziale

$p(x) = \exp(\lambda_0 + \sum_i \lambda_i g_i(x))/Z$

Esempi classici:

Vincoli	Distribuzione MaxEnt
Supporto [a,b], nessun altro	Uniforme
Reali, media μ, varianza σ²	Normale N(μ,σ²)
Non-negativo, media μ	Esponenziale λ=1/μ
Interi ≥0, media μ	Geometrica

Connessione fisica: In fisica statistica, la distribuzione di Boltzmann è MaxEnt dato vincolo su energia media. Collega probabilità, informazione e termodinamica.

Informazione Mutua

L'informazione mutua quantifica quanto apprendere X riduce incertezza su Y:

$I(X;Y) = H[X] + H[Y] - H[X,Y] = KL(p(x,y)||p(x)p(y))$

Proprietà:

- Simmetrica: $I(X;Y) = I(Y;X)$
- Non-negativa: $I(X;Y) \geq 0$
- Zero $\Leftrightarrow$ X,Y indipendenti

In inferenza: $I(\theta;x)$ misura quanto i dati informano sui parametri:

$I(\theta;x) = H[\theta] - E[H[\theta|x]] = E_x[KL(p(\theta|x)||p(\theta))]$

Progettazione esperimenti: Massimizza $I(\theta;x)$ per scegliere l'esperimento più informativo.

Oltre i Parametri Finiti

L'inferenza tradizionale assume parametri in $\mathbb{R}^k$ (finito-dimensionale). Ma molti problemi richiedono stimare oggetti infinito-dimensionali: funzioni, distribuzioni, curve.

L'inferenza bayesiana non-parametrica estende il framework a spazi funzionali - distribuzioni di probabilità su funzioni.

Processi Gaussiani (GP)

Definizione: Un GP è una distribuzione su funzioni. Formalmente, una collezione $\{f(x)\}_{x\in\mathcal{X}}$ dove qualsiasi sottoinsieme finito ha distribuzione gaussiana congiunta.

Specificazione completa:

- Funzione media: m(x) = E[f(x)]
- Kernel (covarianza): k(x,x') = Cov[f(x),f(x')]

Notazione: f ~ GP(m,k)

Kernel comuni:

Squared Exponential (RBF): $k(x,x') = \sigma^2\exp(-||x-x'||^2/2\ell^2)$

Parametri: σ^2 (varianza), ℓ (lengthscale) Proprietà: Infinitamente differenziabile, smooth

Matérn: $k(x,x') = \sigma^2(2^{1-v}/\Gamma(v))(\sqrt{(2v)}r/\ell)^v K_v(\sqrt{(2v)}r/\ell)$

Parametro v controlla smoothness: $v\to\infty$ converge a RBF

Periodic: $k(x,x') = \sigma^2\exp(-2\sin^2(\pi|x-x'|/p)/\ell^2)$

Per funzioni con periodicità p

Regressione con GP:

Prior: f ~ GP(0,k) Likelihood: y = f(X) + ε, ε ~ N(0,σ□²I) Posteriore: f|y ~ GP(m*,k*)

dove: $m^*(x) = k(x,X)[K+\sigma^2I]^{-1}y$ $k^*(x,x') = k(x,x') - k(x,X)[K+\sigma^2I]^{-1}k(X,x')$

K è matrice Gram: $K_{i\square} = k(X_i,X_\square)$

Interpretazione:

- Lontano da dati: posteriore ≈ prior
- Vicino a dati: posteriore attratto verso osservazioni
- Incertezza alta lontano da dati, bassa vicino

Applicazioni:

- Regressione non-parametrica
- Ottimizzazione bayesiana (acquisizione basata su incertezza)
- Emulazione di simulatori costosi
- Calibrazione di modelli complessi

Processi di Dirichlet (DP)

Definizione: Il DP è una distribuzione di probabilità su distribuzioni. Prior non-parametrico standard per clustering con numero sconosciuto di cluster.

Caratterizzato da:

- Concentrazione $\alpha > 0$
- Base G_0 (distribuzione)

Notazione: $G \sim DP(\alpha, G_0)$

Proprietà:

- $E[G] = G_0$
- $Var[G] \propto 1/\alpha$
- G è discreta con prob 1 (anche se G_0 continua!)

Stick-Breaking Construction:

$\pi_k = \beta_k \prod_{j=1}^{k-1}(1-\beta_j), \beta_k \sim Beta(1,\alpha)\ \theta_k \sim G_0\ G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}$

Visualizza: "spezza un bastone" ripetutamente, assegna pesi ai pezzi.

Chinese Restaurant Process (CRP):

Metafora: Ristorante con infiniti tavoli. Cliente n si siede:

- Al tavolo k occupato con prob $n_k/(n-1+\alpha)$
- A nuovo tavolo con prob $\alpha/(n-1+\alpha)$

dove n_k = numero clienti al tavolo k.

DP Mixture Models:

$y_i|\theta_i \sim F(\theta_i)$ (likelihood) $\theta_i|G \sim G$ (mixture) $G \sim DP(\alpha, G_0)$ (prior su G)

Il DP induce prior su numero cluster che:

- Cresce logaritmicamente con n: $E[K] \approx \alpha \log(1+n/\alpha)$
- Penalizza automaticamente complessità
- Permette cluster nuovi anche dopo aver osservato molti dati

Inferenza: Tipicamente MCMC (Gibbs con CRP) o variational inference.

Algoritmo Gibbs per DP mixture:

Per ogni i:

1. Rimuovi y_i dalla sua classe corrente
2. Campiona nuova classe per y_i:
    - Classe esistente k con prob $\propto n_{-i,k} \times p(y_i|\theta_k)$
    - Nuova classe con prob $\propto \alpha \times \int p(y_i|\theta) dG_0(\theta)$
3. Se nuova classe, campiona $\theta_{new} \sim p(\theta|y_i, G_0)$

Hierarchical Dirichlet Process (HDP)

Estensione del DP per dati raggruppati (es. documenti, topic modeling).

Struttura: $G_0 \sim DP(\gamma, H)$ (base shared) $G_j \sim DP(\alpha, G_0)$ (group-specific) $\theta_{ji} \sim G_j$ (per gruppo j, item i)

HDP permette cluster condivisi tra gruppi ma con proporzioni diverse.

Applicazione classica: Topic modeling: Documenti condividono topics (cluster globali) ma con proporzioni diverse per documento.

Indian Buffet Process (IBP)

Prior non-parametrico per feature allocation: quante feature latenti (potenzialmente infinite) spiegano i dati?

Metafora: Infiniti piatti in buffet. Cliente n:

- Prende piatti già presi da altri con prob $\propto$ popolarità
- Prova Poisson(α/n) nuovi piatti

Risultato: Matrice binaria infinita N×∞ dove Z_{nk}=1 se oggetto n ha feature k.

Proprietà:

- Numero atteso di feature presenti: $E[K_+] \approx \alpha \cdot H_N$ (armonico)
- Sparsità: la maggior parte feature non presente nella maggior parte oggetti

Applicazioni:

- Decomposizione matriciale (Z×W)
- Learning di rappresentazioni
- Network latent features

Dalla Correlazione alla Causazione

L'inferenza statistica tradizionale (inclusa bayesiana standard) riguarda associazioni. Ma scienza e decisioni richiedono comprensione causale: cosa succederebbe se interveniamo?

Il framework di Judea Pearl estende la probabilità bayesiana per gestire la causalità.

Grafi Causali (DAG)

Un grafo diretto aciclico rappresenta assunzioni causali:

- Nodi: variabili
- Archi: X→Y significa "X causa Y"

Modello Strutturale Causale (SCM):

$Y = f_Y(Pa(Y), U_Y)$

Ogni variabile è funzione deterministica di genitori + rumore esogeno.

d-separation: Criterio per indipendenze condizionali:

Chain: $X \to Z \to Y \Rightarrow X \perp Y|Z$ *Fork*: $X \leftarrow Z \to Y \Rightarrow X \perp Y|Z$
Collider: $X \to Z \leftarrow Y \Rightarrow X \perp Y|Z$ (ma $X \perp Y$ marginalmente!)

Il collider è controintuitivo: condizionare su Z induce dipendenza (selection bias).

L'Operatore do()

do(X=x): Rappresenta un intervento che fissa X=x, tagliando tutte le frecce entranti in X nel DAG.

Distinzione fondamentale:

- **P(Y|X=x)**: distribuzione osservazionale - probabilità di Y quando osserviamo X=x
- **P(Y|do(X=x))**: distribuzione interventistica - probabilità di Y quando forziamo X=x

Queste sono generalmente diverse!

Esempio classico: Ice cream sales (I) e drownings (D) sono correlati - entrambi causati da temperatura (T):

$T \to I$ $T \to D$

Osservazionalmente: P(D|I=alto) è alto (correlazione) Causalmente: P(D|do(I=alto)) = P(D) (nessun effetto causale)

Aumentare le vendite del gelato non causa annegamenti. La correlazione è spuria, dovuta al confounder T.

Le Tre Regole del do-Calculus

Pearl sviluppò tre regole per manipolare espressioni con do(), permettendo di computare effetti causali da dati osservazionali sotto condizioni di identificabilità.

Sia G il DAG causale, $G_{\bar{X}}$ il grafo con frecce entranti in X rimosse, $G_{\underline{X}}$ il grafo con frecce uscenti da X rimosse.

Regola 1 (Inserimento/rimozione osservazioni): $P(Y|do(X),Z,W) = P(Y|do(X),W)$ se $(Y \perp Z | X,W)$ in $G_{\bar{X}}$

Osservare Z è irrilevante se Z è indipendente da Y dato X,W nel grafo manipolato.

Regola 2 (Scambio azione/osservazione): $P(Y|do(X),do(Z),W) = P(Y|do(X),Z,W)$ se $(Y \perp Z | X,W)$ in $G_{\{\bar{X},Z\}}$

Intervento su Z può essere sostituito da osservazione se Z non ha frecce entranti che influenzano Y.

Regola 3 (Inserimento/rimozione azioni): P(Y|do(X),do(Z),W) = P(Y|do(X),W) se (Y ⊥ Z | X,W) in G_{X̄,Z(W)}

Intervento su Z è irrilevante sotto certe condizioni.

Teorema di Identificazione: Un effetto causale P(Y|do(X)) è identificabile da dati osservazionali se può essere espresso usando solo distribuzioni osservazionali tramite applicazioni ripetute delle tre regole.

Backdoor e Frontdoor Adjustment

Backdoor Criterion: Un set di variabili Z soddisfa il criterio backdoor per (X→Y) se:

1. Z blocca tutti i percorsi backdoor da X a Y
2. Z non contiene discendenti di X

Se Z soddisfa backdoor, allora: P(Y|do(X=x)) = ∑_z P(Y|X=x,Z=z)P(Z=z)

Esempio: Nel DAG T→I, T→D:

- Z = {T} soddisfa backdoor per I→D
- P(D|do(I)) = ∑_t P(D|I,T=t)P(T=t)

Aggiustando per temperatura, eliminiamo confounding.

Frontdoor Criterion: Più sottile. Se X→M→Y e esiste confounding non osservato U→X e U→Y:

- Non possiamo usare backdoor (U non osservato)
- Ma se M soddisfa frontdoor, possiamo ancora identificare P(Y|do(X))

Formula frontdoor: P(Y|do(X=x)) = ∑*m P(M=m|X=x)* ∑{x'} P(Y|M=m,X=x')P(X=x')

Counterfactuals Bayesiani

Counterfactual: "Cosa sarebbe successo se X fosse stato diverso?"

Notazione: Y_x - valore che Y avrebbe assunto se X fosse stato fissato a x.

Tre livelli della scala causale (Pearl):

1. **Associazione**: P(Y|X) - osservazione passiva
2. **Intervento**: P(Y|do(X)) - manipolazione
3. **Counterfactual**: P(Y_x|Y'=y',X'=x') - retrospezione

Ogni livello richiede assunzioni più forti.

Calcolo counterfactual:

1. **Abduction**: Aggiorna distribuzione su variabili esogene U dato evidenza osservata $P(U|e) \propto P(e|U)P(U)$

2. **Action**: Modifica modello secondo intervento do(X=x)

3. **Prediction**: Calcola $P(Y_x|e)$ usando modello modificato e posteriore su U

Esempio medico: Paziente preso farmaco X=1, outcome Y=1 (guarigione). Sarebbe guarito senza farmaco?

$P(Y_0=1 \mid X=1, Y=1) = ?$

Questo richiede modello SCM completo con assunzioni su meccanismi causali e variabili non osservate.

Inferenza Causale da Dati Osservazionali: Sfide

Problema fondamentale: Non possiamo inferire causalità da correlazione senza assunzioni.

Confounding: Variabili comuni causano sia X che Y. Correlazione X-Y può essere spuria.

Reverse causation: Y causa X invece di viceversa.

Selection bias: Campione non rappresentativo induce associazioni spurie.

Measurement error: Errori sistematici distorcono associazioni.

Approcci:

1. **Randomized Controlled Trials**: Gold standard - randomizzazione elimina confounding
2. **Instrumental Variables**: Trova variabile Z che influenza X ma non Y direttamente
3. **Regression Discontinuity**: Sfrutta discontinuità nel trattamento
4. **Difference-in-Differences**: Confronta cambiamenti nel tempo
5. **Propensity Score Matching**: Bilancia covariate tra trattati e controlli

Framework bayesiano per causalità: Possiamo porre prior su strutture causali (DAG), stimare posteriori su grafi e parametri simultaneamente. Bayesian Causal Networks integrano incertezza strutturale e parametrica.

Il Paradosso di Bertrand

Bertrand (1889) mostrò che il "principio di indifferenza" porta a contraddizioni.

Problema: Corda casuale su cerchio. Probabilità che sia più lunga del lato di triangolo equilatero inscritto?

Tre interpretazioni di "casuale":

1. **Angolo random**: Endpoint fisso, angolo uniforme → P = 1/3
2. **Punto medio random**: Punto medio uniforme su diametro → P = 1/2
3. **Area random**: Punto medio uniforme su disco → P = 1/4

Stessi dati, interpretazioni diverse del prior "uniforme" → risposte diverse!

Lezione: Non esiste prior univocamente "non-informativo". La scelta dipende dalla parametrizzazione, che riflette assunzioni sostantive.

Il Paradosso di Borel

Setup: Parametro $\theta \in [0,1]$. Prior uniforme $p(\theta) = 1$.

Riparametrizza: $\varphi = \theta/(1-\theta)$, $\varphi \in [0,\infty)$.

Trasformazione: $p(\varphi) = p(\theta)|d\theta/d\varphi| = 1/(1+\varphi)^2$

Questo non è uniforme! Il prior "non-informativo" su θ induce prior informativo su φ.

Implicazione: L'uniformità non è invariante. Quale parametrizzazione è "naturale"? Dipende dal problema.

Soluzione parziale: Prior di Jeffreys è invariante, ma può essere improprio o portare ad altri problemi.

Il Problema delle Prefiche (Marginalization Paradox)

Setup: Due parametri (θ_1,θ_2). Prior di Jeffreys sulla congiunzione:

$p(\theta_1,\theta_2) \propto \sqrt{\det(I(\theta_1,\theta_2))}$

Marginalizzando su θ_2: $p(\theta_1) = \int p(\theta_1,\theta_2)d\theta_2$

Questo generalmente NON è uguale al prior di Jeffreys calcolato direttamente per θ_1!

Paradosso: Il prior "oggettivo" dipende da quali parametri includiamo nel modello, anche se non sono di interesse.

Posizione moderna: Non esiste prior completamente oggettivo. Ogni scelta riflette assunzioni. La trasparenza è più importante della pretesa oggettività.

Il Paradosso di Lindley

Setup: Test di ipotesi H_0: $\theta=\theta_0$ vs H_1: $\theta\neq\theta_0$.

Con campioni molto grandi:

- **Test frequentista**: Rigetta H_0 con p-value minuscolo
- **Analisi bayesiana**: Alta probabilità posteriore per H_0

Come può essere?

Spiegazione: I due approcci rispondono a domande diverse.

Frequentista calcola: P(dati estremi come questi | H_0 vera) Bayesiano calcola: P(H_0 vera | dati)

Per n grande, anche deviazioni minuscole da H_0 diventano statisticamente rilevabili (p-value piccolo), ma se l'effetto è molto piccolo, il modello H_0 può comunque essere la migliore approssimazione (posteriore alta).

Esempio: Test se moneta è equa. Con 1 milione lanci: 500,100 teste.

- Frequentista: $p < 0.0001$, rigetta fortemente equità
- Bayesiano: P(equa) ancora alta - differenza da 0.5 è minuscola (0.5001)

La Curse of Dimensionality in MCMC

In alta dimensionalità (k grande), MCMC affronta problemi severi:

Problema 1 - Volume esponenziale: Lo spazio cresce come R^k. Per k=100 e R=10, ci sono 10^100 celle - più di atomi nell'universo.

Problema 2 - Concentrazione: In alta dimensione, quasi tutta la massa è in un guscio sottile, non al centro né alle code. Il "modo tipico" non è il modo (massimo).

Distanza di un punto random dall'origine in sfera k-dimensionale:

- k=2: distribuzione spread
- k=100: concentrata in guscio stretto

Problema 3 - Mixing lento: Random walk Metropolis richiede $O(k^2)$ iterazioni per esplorare. HMC migliora a O(k^{5/4}) ma ancora cresce.

Problema 4 - Diagnostica: È difficile sapere se abbiamo esplorato adequatamente. R-hat e ESS possono dare falsi positivi.

Soluzioni parziali:

- HMC/NUTS (sfrutta gradienti)
- Riparametrizzazione intelligente
- Variational inference (approssimazione più veloce)
- Reduced dimensionality via PCA o altri

Il Problema della Misspecificazione

Realtà: Il modello è sempre sbagliato. "All models are wrong, some are useful" (Box).

Domanda: Cosa fa l'inferenza bayesiana sotto misspecificazione?

Risposta: Il posteriore converge al valore che minimizza KL dal vero processo:

$\theta^* = \text{argmin}_\theta \, KL(p_{true} \,||\, p_{model}(\cdot|\theta))$

È la "best approximation" nella famiglia specificata.

Problema: Gli intervalli di credibilità possono sottostimare drasticamente l'incertezza vera perché non tengono conto dell'incertezza del modello stesso.

Approcci:

1. **Model averaging**: Porre prior su spazio di modelli, mediare
2. **Robust priors**: Prior con code pesanti
3. **Sensitivity analysis**: Testare multiple specificazioni
4. **Cross-validation**: Valutazione predittiva out-of-sample

Il Problema dei Prior "Convenient" vs "Realistic"

Tensione: Prior coniugati rendono calcoli analitici, ma raramente riflettono credenze reali.

Esempio: Beta prior per probabilità. Beta(1,1) è uniforme (conveniente), ma implica che p=0.001 e p=0.999 sono altrettanto plausibili di p=0.5. È realistico?

Per fenomeni biologici, sociali, un prior che concentra massa vicino a 0.5 sarebbe più realistico, ma meno trattabile.

Trade-off moderno: Con MCMC, possiamo usare prior realistici anche non-coniugati. Ma c'è ancora tentazione di scegliere prior per convenienza computazionale.

Raccomandazione: Esplicitare quando prior è chosen per convenienza. Fare sensitivity analysis.

Bayesian Deep Learning

Problema: Neural networks hanno milioni di parametri. Inferenza bayesiana piena è intrattabile.

Approcci:

1. Dropout come Approssimazione Bayesiana (Gal & Ghahramani):

- Dropout durante training E test
- Ensemble di reti ≈ approssimazione variazionale
- Fornisce barre d'errore su predizioni

2. Variational Inference per NN:

- Approssima posteriore su pesi con $q(w|\varphi)$
- Minimizza $KL(q||p)$
- Bayes by Backprop algorithm

3. Laplace Approximation:

- Approssima posteriore con Gaussiana centrata su MAP
- Hessiana alla moda → matrice covarianza
- Scalabile con modern automatic differentiation

4. Stochastic Weight Averaging Gaussian (SWAG):

- Approssima posteriore da trajectory di SGD
- Cattura geometria locale attorno a ottimi

Sfide:

- Scalabilità: miliardi di parametri
- Multimodalità: molti ottimi locali
- Cold posterior: posteriori empirici troppo concentrati
- Calibrazione: incertezze bayesiane non sempre ben calibrate

Approximate Bayesian Computation (ABC)

Problema: Molti modelli scientifici hanno simulatori complessi ma nessuna likelihood esplicita.

Esempi:

- Modelli evolutivi in biologia
- Climate models
- Agent-based models in economia

Idea ABC: Anziché valutare p(dati|θ), simula:

1. Campiona θ ~ p(θ)
2. Simula dati finti x* ~ p(·|θ)
3. Accetta θ se d(x*, x_obs) < ε

dove d(·,·) è una distanza e ε è tolleranza.

Giustificazione: Approssima P(θ | d(x*,x_obs) < ε) ≈ P(θ|x_obs)

per ε piccolo.

Varianti:

ABC-MCMC: Usa logica ABC in passo accettazione Metropolis

ABC-SMC: Sequential Monte Carlo con tolleranze decrescenti

Regression adjustment: Usa regressione per correggere bias da ε>0

Neural density estimation: Usa neural networks per stimare likelihood ratio

Sfide:

- Scelta di distanza d e statistics
- High dimensionality dei dati
- Computational cost
- Determinare ε appropriato

Bayesian Optimization

Problema: Ottimizzare f(x) che è costosa da valutare (es. simulazioni, esperimenti reali) e non abbiamo gradients.

Approccio bayesiano:

1. Porre GP prior su f
2. Valutare f in pochi punti → posteriore GP
3. Usare posteriore per decidere dove valutare next
4. Iterate

Acquisition Functions: Bilanciano exploration vs exploitation

Expected Improvement (EI): $EI(x) = E[\max(f(x) - f_{best}, 0) \mid dati]$

Upper Confidence Bound (UCB): $UCB(x) = \mu(x) + \kappa\sigma(x)$

dove μ,σ da GP posteriore, κ controlla exploration.

Knowledge Gradient: Valore dell'informazione per decisione futura ottima.

Applicazioni:

- Hyperparameter tuning in ML
- Experimental design
- Drug discovery
- Engineering design optimization

Estensioni:

- Multi-objective optimization
- Constrained optimization
- Batch bayesian optimization
- High-dimensional BO (via embeddings, random projections)

Bayesian Causal Discovery

Problema: Inferire struttura causale (DAG) dai dati, non solo parametri.

Approccio: Porre prior su spazio dei DAG, computare posteriore.

Challenges:

- Spazio di DAG è super-esponenziale: k nodi → ~k!·2^{k(k-1)/2} DAG
- Equivalenza di Markov: DAG diversi possono implicare stesse indipendenze
- Interventions aiutano ma sono costosi

Metodi:

Score-based: Assegna score a ciascun DAG (es. BIC, BDe), cerca massimo

Constraint-based: Testa indipendenze condizionali, elimina edges inconsistenti

Hybrid: Combina score e constraints

Bayesian: Prior su DAG (uniformi, o preferenza per sparsità), MCMC su spazio DAG

Identifiability results:

- Solo da dati osservazionali: Markov equivalence class
- Con interventions: Può identificare orientamento esatto
- Con nonlinear additive noise models: Può identificare anche senza interventions

Bayesian Inverse Problems

Problema: Data osservazione indiretta $y = G(x) + \varepsilon$, inferire x.

Esempi:

- Medical imaging (CT, MRI): Da proiezioni → 3D structure
- Seismology: Da onde → struttura Terra
- Astronomy: Da immagini → proprietà oggetti celesti

Framework bayesiano:

Prior: $p(x)$ - regolarizzazione, smoothness Likelihood: $p(y|x)$ - modello di rumore Posteriore: $p(x|y) \propto p(y|x)p(x)$

Vantaggi:

- Quantificazione incertezza
- Incorporazione di conoscenza pregressa (prior)
- Handling ill-posedness

Metodi computazionali:

- MCMC (costoso in high-dim)
- Variational inference
- Laplace approximation
- Ensemble methods

Probabilistic Programming

Idea: Linguaggi dove programmi definiscono modelli probabilistici. Inference automatica.

Esempi di linguaggi:

Stan: HMC/NUTS, ottimo per continuous parameters **PyMC**: Python, flessibile **Turing.jl**: Julia, composable **Pyro/NumPyro**: Python, basato su JAX, scalabile **TensorFlow Probability**: Integrato con TF/Keras

Workflow tipico:

model = define_model(prior, likelihood)

posterior = infer(model, data, method=NUTS)

predictions = predict(posterior, new_data)

Vantaggi:

- Democratizza bayesian inference
- Separa modeling da inference
- Diagnostics e visualization automatiche

Sfide:

- Debugging è difficile (probabilistic)
- Performance può essere subottimale vs hand-coded
- Richiede comprensione di concepts sottostanti

Scalable Bayesian Inference per Big Data

Problema: Dataset con miliardi di punti. Likelihood evaluation O(n) per iterazione MCMC è proibitivo.

Approcci:

1. Stochastic Gradient MCMC:

- Usa mini-batch per stimare gradients
- Add strategical noise per mantenere convergenza corretta
- SGLD, SGHMC varianti

2. Variational Inference:

- Ottimizzazione vs sampling
- Naturalmente batch-based
- Scales meglio di MCMC

3. Consensus Monte Carlo:

- Parallelizza: dividi dati in shards
- Inference indipendente per shard
- Combina posteriori (product of experts, consensus)

4. Subsampling with pseudo-marginal:

- Usa subset di dati per stimare likelihood
- Correggi bias via importance weighting

5. Inducing points methods (per GP):

- Approssima GP con punti induttivi finiti
- Scalabilità da $O(n^3)$ a $O(m^2n)$ con $m << n$

Bayesian Reinforcement Learning

Problema: Agente interagisce con ambiente. Come apprendere e agire ottimamente sotto incertezza?

Framework:

- Stati S, Azioni A, Rewards R
- Dinamiche sconosciute: $P(s'|s,a)$, $R(s,a)$
- Prior su dinamiche e reward

Approccio bayesiano:

1. Mantieni posteriore su MDP
2. Aggiorna posteriore con esperienza
3. Agisci per massimizzare expected reward rispetto a posteriore

Information gain: Azioni che riducono incertezza hanno valore intrinseco (exploration).

Metodi:

Thompson Sampling: Campiona parametri da posteriore, agisci ottimamente per quel campione

Bayes-Adaptive MDP: Stato aumentato include belief state. Solvi MDP espanso.

Upper Confidence Bound: Usa intervalli credibilità per bilanciare exploration-exploitation

Sfide:

- Computational intractability di bellman equation bayesiano
- Belief state è high-dimensional
- Approssimazioni necessarie (particles, moments)

L'Unificazione di Induzione e Probabilità

Il teorema di Bayes realizza il sogno di Leibniz: una "calculus ratiocinator" per il ragionamento induttivo. Fornisce regole precise per come evidenza dovrebbe modificare credenza.

Prima di Bayes: induzione era informale, guidata da intuizione Dopo Bayes: induzione diventa mathematizzabile, verificabile, replicabile

Ma: Questo riduce davvero induzione a deduzione probabilistica? Hume's problem of induction rimane: perché il futuro dovrebbe assomigliare al passato? Il teorema di Bayes assume che c'è un θ stabile da apprendere - assume già regolarità.

Oggettività vs Soggettività: Una Falsa Dicotomia?

Il dibattito frequentismo-bayesianesimo è spesso inquadrato come oggettività vs soggettività.

Ma: Entrambi hanno elementi soggettivi e oggettivi:

- Frequentismo: scelta di test statistic, livello α - soggettivi
- Bayesianesimo: prior - soggettivo; likelihood e aggiornamento - oggettivi

Proposta: Abbandonare dicotomia. Focus su:

- Trasparenza: esplicitare assunzioni
- Robustezza: sensitivity analysis
- Coerenza: seguire principi logici (assiomi probabilità)

Conclusione

Il teorema di Bayes, da una semplice formula derivata nel XVIII secolo, è diventato uno dei pilastri concettuali e computazionali della scienza moderna. La sua influenza permea:

- **Statistica e machine learning:** Fondamento dell'inferenza probabilistica
- **Scienza dei dati:** Quantificazione dell'incertezza e model selection
- **Intelligenza artificiale:** Reasoning under uncertainty, agenti razionali
- **Medicina:** Diagnosi, screening, trial clinici
- **Fisica:** Quantistica, cosmologia, analisi dati sperimentali
- **Economia e finanza:** Decision theory, risk management
- **Neuroscienze:** Modelli di percezione e cognizione
- **Filosofia:** Epistemologia, confirmazione scientifica

La forza del teorema di Bayes risiede nella sua capacità di:

1. Formalizzare il ragionamento razionale in condizioni di incertezza
2. Invertire la direzione causale (da effetti a cause)
3. Aggiornare credenze in modo coerente con nuova evidenza
4. Quantificare esplicitamente l'incertezza
5. Fornire un framework unificato per inferenza e decisione